Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Graph-Sparse LDA: A Topic Model with Structured Sparsity

Finale Doshi-Velez
Harvard University
Cambridge, MA 02138
finale@seas.harvard.edu

Byron C Wallace
University of Texas at Austin
Austin, TX 78701
byron.wallace@utexas.edu

Ryan Adams
Harvard University
Cambridge, MA 02138
rpa@seas.harvard.edu

Abstract

Topic modeling is a powerful tool for uncovering latent structure in many domains, including medicine, finance, and vision. The goals for the model vary depending on the application: sometimes the discovered topics are used for prediction or another downstream task. In other cases, the content of the topic may be of intrinsic scientific interest. Unfortunately, even when one uses modern sparse techniques, discovered topics are often difficult to interpret due to the high dimensionality of the underlying space. To improve topic interpretability, we introduce Graph-Sparse LDA, a hierarchical topic model that uses knowledge of relationships between words (e.g., as encoded by an ontology). In our model, topics are summarized by a few latent *concept-words* from the underlying graph that explain the observed words. Graph-Sparse LDA recovers sparse, interpretable summaries on two real-world biomedical datasets while matching state-of-the-art prediction performance.

Introduction

Probabilistic topic models (Blei, Ng, and Jordan 2003; Steyvers and Griffiths 2007) were originally developed to discover latent structure in unorganized text corpora. However these models provide a powerful general framework for uncovering structure in data drawn from many domains (e.g., medicine, finance and vision, to name a few). In the popular Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) model, *topics* are distributions over the words in a vocabulary, and documents are summarized by the mixture of topics that they contain. Here, a "word" is anything that can be counted and a "document" is an observation. LDA has been applied to a diverse set of tasks, including finding scientific topics in articles (Griffiths and Steyvers 2004), classifying images (Fei-Fei and Perona 2005), and recognizing human actions (Wang and Mori 2009). The modeling objective varies depending on the application: in some cases, topic models are used to provide compact summaries of documents that can then be used for downstream tasks such as prediction, classification, or recognition. In other situations, the discovered topics themselves may be of independent interest. For example, a clinician may want to understand *why* a certain topic within their patient's data is correlated with mortality (Ghassemi et al. 2012).

Applications in which interpretation is paramount present unique challenges and opportunities for topic modeling. Typically, topics are distributions over the words in a (very large) vocabulary that is usually assumed to be unstructured, i.e., words do not have *a priori* relationships. Sparse topic models, e.g., (Archambeau, Lakshminarayanan, and Bouchard 2011; Williamson et al. 2010), offer some interpretability via the constraint that many of a topic's word probabilities should be zero. Unfortunately, when vocabularies are large, there may still be hundreds of words with non-zero probabilities. Enforcing sparsity alone is therefore not sufficient to induce interpretable topics.

We propose a new strategy for achieving interpretability: exploiting *controlled structured vocabularies*, which exist in many technical domains. These structures encode known relationships between words. For example, diseases are organized into billing hierarchies, and clinical concepts are related by directed acyclic graphs (DAGs) (Bodenreider 2004). Keywords for biomedical publications are organized in a hierarchy known as MeSH (Lipscomb 2000). Genes are organized into pathways and interaction networks. Such structures often summarize large bodies of scientific research and human thought. While these structured vocabularies are necessarily imperfect, they have the important property that they (by definition) represent how *domain experts codify knowledge*, and thus might help to create models that such experts can meaningfully use and interpret. And because they were designed to be understood by humans, these relationships provide a form of information unique from any learned ontology.

Existing topic modeling machinery is not equipped to leverage controlled structured vocabularies. We propose Graph-Sparse LDA (GS-LDA), a new model that exploits DAG-structured vocabularies to induce interpretable topics that still summarize the data well. Our approach is appropriate when documents are annotated with structured vocabulary terms, e.g., biomedical articles with MeSH headers, genes with known interactions, and species with known taxonomies. GS-LDA introduces an additional layer of hierarchy into the standard LDA model: instead of topics being distributions over observed words, they are distributions over *concept-words*, which then generate observed words

via a noise process that is informed by the structure of the vocabulary (see example in Figure 1). By exploiting the structure in the vocabulary to induce sparsity, we recover topics that are more interpretable to domain experts.

We demonstrate GS-LDA on two real-world applications. The first is a collection of diagnoses for patients with autism spectrum disorder. For this we use a diagnosis hierarchy (Bodenreider 2004) to recover clinically relevant subtypes described by a small set of concepts. The second is a corpus of biomedical abstracts annotated with hierarchically-structured Medical Subject Headings (MeSH) (Lipscomb 2000). Here GS-LDA identifies meaningful, concise groupings (topics) of MeSH terms for use in biomedical literature retrieval tasks. In both cases, the topic models found by GS-LDA have the same or better predictive performance as a state-of-the-art sparse topic model (Latent IBP compound Dirichlet Allocation (Archambeau, Lakshminarayanan, and Bouchard 2011)) while providing much sparser topic descriptions. To efficiently sample from this model, we introduce a novel inference procedure that prefers moves along manifolds of constant likelihood to identify sparse solutions.

Graph-Sparse LDA

In this paper, our data are documents that are modeled using the standard "bag of words" representation. Let the data X consist of the counts of each of the V words in the vocabulary for each of the N documents. The standard LDA model (Blei, Ng, and Jordan 2003) posits the following generative process for the words w_{in} comprising each document (data instance) in X:

$$B_n \sim \text{Dirichlet}(\alpha_B 1_K) \quad (1)$$
$$A_k \sim \text{Dirichlet}(\alpha_A 1_V) \quad (2)$$
$$z_{in} \mid B_n \sim \text{Discrete}(B_n) \quad (3)$$
$$w_{in} \mid z_{in}, \{A_k\} \sim \text{Discrete}(A_{z_{in}}) \quad (4)$$

where K is the number of topics. The rows of the $N \times K$ matrix B are the document-specific distributions over topics, and the $K \times V$ matrix A represents each topic's distribution over words. The notation A_k refers to the k^{th} row of A. The z_{in} encode the topic to which the i^{th} word in document n was assigned, and $w_{in} \in 1, \ldots, V$ is the i^{th} word in document n.

Our Bayesian nonparametric model, GS-LDA, builds upon a recent nonparametric extension of LDA, Latent IBP compound Dirichlet Allocation (LIDA) (Archambeau, Lakshminarayanan, and Bouchard 2011). LIDA introduces sparsity over both the document-topic matrix B and the topic-word matrix A using a three-parameter Indian Buffet Process. This prior encodes a preference for describing each document with a few topics and each topic with a few words. We extend LIDA by assuming that words in our document have known relationships that form a tree or DAG, and that nearby groups of terms—as defined with respect to the graph structure—are associated with specific phenomena. For example, in a biomedical ontology, nodes on one sub-tree may correspond to a particular virus (e.g., HIV) and a different sub-tree may describe a specific drug or treatment (e.g., anti-retrovirals) used to treat HIV. Papers investigating anti-retrovirals for treatment of HIV would then tend to have terms drawn from both sub-trees. Intuitively, we would like to uncover these sub-trees as the concepts underpinning a topic.

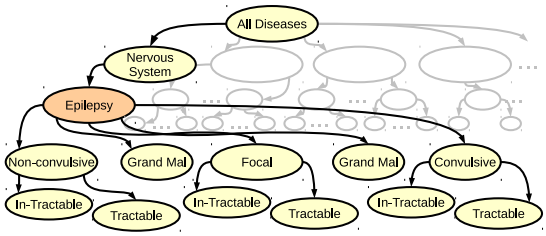


Figure 1: Simplified section of the ICD9-CM diagnostic hierarchy. Here, "Epilepsy" might be a good *concept-word* to summarize the very specific forms of epilepsy that are its descendants. Knowing that a patient has epilepsy may also explain instances of "Central Nervous System Disorder" or even "Disease."

Using concept-words to summarize the words in a topic is natural in many scenarios because structured vocabularies are often both very specific and inconsistently applied. For example, a trial may be annotated with the term *antiviral agents* or its child *anti-retroviral agents*. Nearby words in the vocabulary can be thought of as having been generated from the same core concept. Our model posits that a topic is made up of a sparse set of concept-words that can explain words that are its ancestors or descendants (Figure 1). Formally, we define the following generative process that introduces $\tilde{w}_{in}$ as the concept word behind observed word w_{in}:

$$\pi_k \sim \text{IBP-Stick}(\gamma_B) \quad (5)$$
$$\rho_v \sim \text{Beta}(\gamma_A / V, 1) \quad (6)$$
$$\bar{B}_{nk} \mid \pi_k \sim \text{Bernoulli}(\pi_k) \quad (7)$$
$$\bar{A}_{kv} \mid \rho_v \sim \text{Bernoulli}(\rho_v) \quad (8)$$
$$B_n \mid \bar{B}_n \sim \text{Dirichlet}(\bar{B}_n \odot \alpha_B 1_K) \quad (9)$$
$$A_k \mid \bar{A}_k \sim \text{Dirichlet}(\bar{A}_k \odot \alpha_A 1_V) \quad (10)$$
$$z_{in} \mid B_n \sim \text{Discrete}(B_n) \quad (11)$$
$$\tilde{w}_{in} \mid z_{in}, \{A_k\} \sim \text{Discrete}(A_{z_{in}}) \quad (12)$$
$$P_v \sim \text{Dirichlet}(\mathcal{O}_v \odot \alpha_P 1_V) \quad (13)$$
$$w_{in} \mid \tilde{w}_{in}, P \sim \text{Discrete}(P_{\tilde{w}_{in}}) \quad (14)$$

where $\odot$ is the element-wise Hadamard product and IBP is the Indian Buffet Process (Griffiths and Ghahramani 2011). We assume both concept words and observed words come from the same vocabulary. As in the standard LDA model, the document-topic matrix B represents the distribution of topics in each document. However, B_n is now masked according to a document-specific vector $\bar{B}_n$, which is the n^{th} row of a matrix $\bar{B}$ that is itself drawn from an IBP with concentration parameter γ_B. Thus $\bar{B}_{nk}$ is 1 if topic k has nonzero probability in document n and 0 otherwise. Similarly, the topic-concept matrix A and the binary topic-concept mask matrix $\bar{A}$ represent the topic matrix and its sparsity pattern, except that now A and $\bar{A}$ represent the

relationship between topics and concept-words. The priors over the document-topic and topic-concept matrices B and A (and their respective masks $\bar{B}$ and $\bar{A}$) follow those in LIDA (Archambeau, Lakshminarayanan, and Bouchard 2011).

The concept-word matrix P describes distributions over words for each concept. The form of the ontology $\mathcal{O}$ determines the sparsity pattern of P: we denote by $\mathcal{O}_w$ a binary vector of length V that is 1 if the concept-word $\tilde{w}$ is a descendant or ancestor of observed word w and 0 otherwise. We illustrate these sparsity constraints in Figure 2, where the dark-shaded concept nodes 1, 2, and 3 can each only explain themselves, and words that are ancestors or descendants. The brown and green nodes are ancestor observed words that are shared by more than one concept word.

Intuitively, the concept-word matrix P allows for variation in the process of domain experts assigning terms to documents (citations, diagnoses, etc.). For example, if a document is about *anti-retroviral agents*, an annotator may describe the document with a key-word nearby in the vocabulary, such as *antiviral agents*, rather than the more specific term. Similarly, a primary care physician using the hierarchy in Figure 1 may note that a patient has *epilepsy* since she is not an expert in neurological disorders, while a specialist might assign the more specific term *Convulsive Epilepsy, Intractable*. More generally, the concept-word matrix P can be thought of as describing a neighborhood of words that could be covered by the same concept. Introducing this additional layer of hierarchy allows us to induce sparse topic-concept matrices A that still explain a large number of observed words. (Note that setting $P = I_V$ recovers LIDA from GS-LDA; GS-LDA is therefore a generalization of LIDA that allows for more structure.)

Inference

In the supplementary materials, we derive a blocked-Gibbs sampler for B, $\bar{B}$, A, $\bar{A}$, and P (as well as for adding and deleting topics). However, Gibbs sampling alone does not give us sparsity in the topic-concept word matrix A fast enough. Mixing is slow because the only time the blocked-Gibbs sampler sets $\bar{A}_{k\tilde{w}} = 0$ is when no counts of $\tilde{w}$ are assigned to topic k *across any of the documents*. When there are many documents, reaching zero counts is unlikely, and the sampler is slow to sparsify the topic-concept word matrix A.[1]

We introduce an MH procedure to encourage moves of the topic concept-word matrix A in directions of greater sparsity through joint moves on both A and P. Given a proposal distribution $Q(A', P' \mid A, P)$, the acceptance ratio for an MH procedure is given by

$$a_{MH} = 1 \wedge \frac{p(X \mid B, A', P')\, p(A')\, p(P')\, Q(A, P \mid A', P')}{p(X \mid B, A, P)\, p(A)\, p(P)\, Q(A', P' \mid A, P)}$$

The prior A prefers sparse topic-concept word matrices A'. However, the likelihood term $p(X \mid B, A, P)$ will generally dominate the prior terms $p(A)$ and $p(P)$.

[1] We focus on A in this section because we found that $\bar{B}$ is faster to mix; each document may not have many words. However, a similar approach could be used to sparsify B.

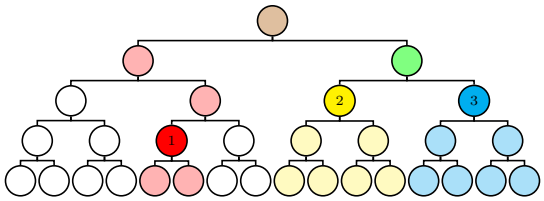


Figure 2: Example tree. Each node (including interior nodes) represents a vocabulary word. A concept-word can explain instances of its descendants and ancestors, e.g., if node 1 is a concept-word, the matrix P would only have non-zero values the nodes in red and brown.

To allow for moves toward greater sparsity, our MH proposal uses two core ideas. First, we use the form of the ontology to propose intelligent split-merge moves for A'. Second, we attempt to make a move that *keeps the likelihood as constant as possible* by proposing a P' such that $AP = A'P'$. Thus, the prior terms $p(A)$ and $p(P)$ will have a larger influence on the move. The form of $Q(A', P' \mid A, P)$ is as follows:

- $Q(A' \mid A, P)$: We choose a random topic k and concept word $\tilde{w}$. Let $D_{\tilde{w}}$ denote the set of concept words that are descendants of $\tilde{w}$ (including $\tilde{w}$). With probability p_{split}, we sample a random vector r from Dirichlet$(1_{|D_{\tilde{w}}|})$ and create a new A'_k with $A'_{k\tilde{w}} = 0$ and $A'_{k\tilde{w}'} = A_{k\tilde{w}'} + rA_{k\tilde{w}}$, $\forall \tilde{w}' \in D_{\tilde{w}}$. Otherwise, we perform the merge $A'_{k\tilde{w}'} = 0$, $\forall \tilde{w}' \in D_{\tilde{w}}$, and $A'_{k\tilde{w}} = \sum_{\tilde{w}' \in D_{\tilde{w}}} A_{k\tilde{w}'}$. This split-merge move corresponds to adjusting probabilities in a sub-graph of the ontology, with the merge move corresponding to moving all the mass to a single node.
- $Q(P' \mid A', A, P)$: Let $P^\star$ be the solution to the optimization problem $\min_{\hat{P}} ||AP - A'\hat{P}||_F^2$, where F denotes the Frobenius norm, with the constraints that each row of $P^\star$ must lie on the simplex and respect the ontology $\mathcal{O}$. This optimization can be solved as a quadratic program with linear constraints. We then sample each row of the proposal P' according to $P'_v \sim \text{Dirichlet}(\beta_{MH} P^\star_v)$. We find in practice that β_{MH} generally needs to be large in order to propose appropriately conservative moves.

While this procedure can still propose moves over the entire parameter space (thus guaranteeing Harris recurrence on the appropriate stationary distribution corresponding to the prior), it guarantees visits to sparse, high-likelihood solutions with high probability.

Results

We demonstrate that our Graph-Sparse LDA model finds interpretable, predictive topics on one toy example and two real-world examples from biomedical domains. In each case we compare our model with the state-of-the-art Bayesian nonparametric topic modeling approach LIDA (Archambeau, Lakshminarayanan, and Bouchard 2011). We focus on LIDA because it subsumes two other popular sparse topic models, the focused topic model (Williamson et al. 2010)

and sparse topic model (Wang and Blei 2009), and because the proposed model is a generalization of LIDA.

We ran all samplers for 250 iterations. To reduce burn-in, The product AP was initialized using an LDA tensor decomposition (Anandkumar et al. 2012) and then factored into A and P using alternating minimization to find a sparse A that enforced the simplex and ontology constraints. A random 1% of each data-set was held out to compute predictive log-likelihoods.

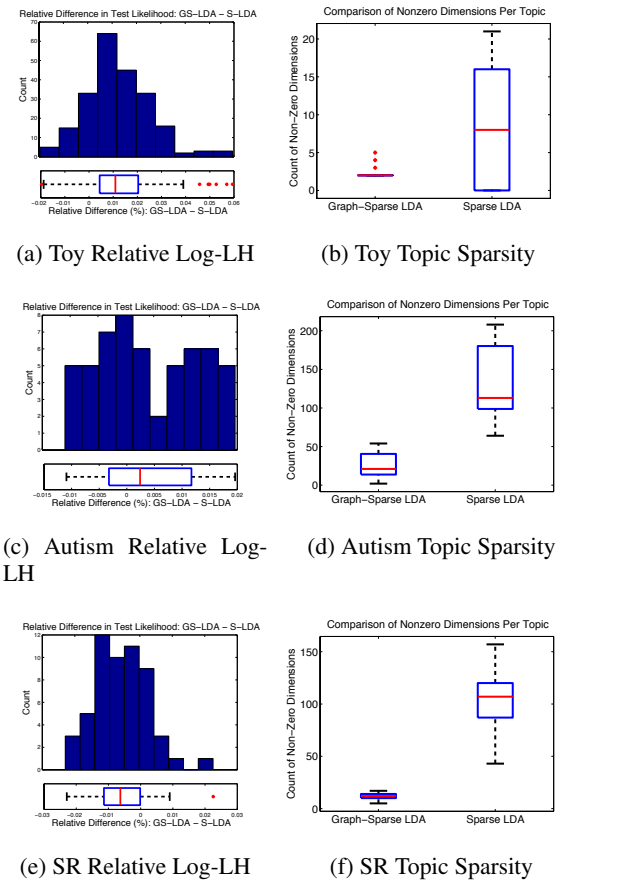
Demonstration on a Toy Problem We first considered a toy problem with a 31-word vocabulary arranged in a binary tree (see Figure 2). There were three underlying topics, each with only a single concept (the three darker nodes in Figure 2, labeled 1, 2, and 3). Each row in the matrix $P_{\tilde{w}}$ uniformly distributed 10% of its probability mass to the ancestors of each concept word and 90% of its probability mass to the concept word's descendants (including itself). Each initialization of the problem had a randomly generated document-topic matrix comprising 1000 documents.

Figures 3a and 3b show the difference in the held-out test likelihoods for the final 50 samples over 20 independent instantiations of the toy problem. The difference in held-out test likelihoods is skewed positive, implying that GS-LDA makes somewhat better predictions than LIDA. More importantly, GS-LDA also recovers a much sparser matrix A, as can be seen in Figure 3b. Of course, that GS-LDA has an additional layer of structure that allows for a very sparse topic concept-word matrix A; LIDA does not have access to the ontology information $\mathcal{O}$. The important point is that by incorporating this available controlled structured vocabulary into our model, we find a solution with similar or better predictive performance than state-of-the-art models with the additional benefit of a much more interpretable structure.

Real World Application: Patterns of Co-Occurring Diagnoses in Autism Spectrum Disorder Autism Spectrum Disorder (ASD) is a complex, heterogenous disease that is often accompanied by many co-occurring conditions such as epilepsy and intellectual disability. We consider a set of 3804 patients with 3626 different diagnoses where the datum X_{nw} corresponds to the number of times patient n received diagnosis w during the first 15 years of life.[2] Diagnoses are organized in a tree-structured hierarchy known as ICD-9CM (Bodenreider 2004). Diagnoses higher up in the hierarchy are less specific (such as "Diseases of the Central Nervous System" or "Epilepsy with Recurrent Seizures," as opposed to "Epilepsy, Unspecified, without mention of intractable epilepsy"). Clinicians may encode a diagnosis at any level of the hierarchy, including less specific ones.

Figure 3c shows the difference in test log-likelihood between GS-LDA and LIDA over 5 independent runs, divided by the overall mean test-likelihood. GS-LDA has slightly better predictive performance—certainly on par with current state-of-the-art topic modeling. However, the use of the ontology results in much sparser topics, as seen in figure 3d. In this application, the topics correspond to possible subtypes in ASD. Being able to concisely summarize these subtypes is an important step for guiding future clinical research.

[2]The Internal Review Board of the Harvard Medical School approved this study.



(a) Toy Relative Log-LH (b) Toy Topic Sparsity

(c) Autism Relative Log-LH (d) Autism Topic Sparsity

(e) SR Relative Log-LH (f) SR Topic Sparsity

Figure 3: The top row shows the difference in held-out test log-likelihoods between GS-LDA and Sparse LDA, divided by the overall mean held-out log-likelihood of both models after burn-in. In three domains, the predictive performance of GS-LDA is within a few percent of LIDA. The second row shows the number of non-zero dimensions in the topic-concept word and the topic-word for GS-LDA and LIDA models, respectively. Results are shown over 20 independent instantiations of the toy problem and 5 independent MCMC runs of the Autism and systematic review (SR) problems.

Finally, Table 1 shows an example of one topic recovered by GS-LDA and its corresponding topic discovered by LIDA. While the corresponding topic in LIDA has very similar diagnoses, using the hierarchy allows for GS-LDA to summarize most of the probability mass in this topic in 6 concept words rather than 119 words. This topic—which shows a connection between the more severe form of ASD, intellectual disability, and epilepsy—as well as the other topics, matched recently published clinical results on ASD subtypes (Doshi-Velez, Ge, and Kohane 2013).

Real World Application: Medical Subject Headings for Biomedical Literature The National Library of Medicine maintains a controlled structured vocabulary of Medical

Table 1: A sample topic from the ASD data. GS-LDA required only 6 concepts to summarize most of probability mass in the topic, while LIDA required 119. For LIDA, we do not show all of the diagnoses associated with the topic, only a sample of the diagnoses summarized by the shown concept words.

Graph-Sparse LDA (6 total nonzero)	LIDA (119 total nonzero)
0.333: Autistic disorder, current or active state	**(1)** *0.213*: Autistic disorder, current or active state
0.203: Epilepsy and recurrent seizures	**(15)**, including *0.052*: Epilepsy, unspecified, without mention of intractable epilepsy, *0.0283*: Localization-related epilepsy and epileptic syndromes with com, *0.023*: Generalized convulsive epilepsy, without mention of intractable epilepsy, *0.008*: Localization-related epilepsy and epileptic syndromes with sim, *0.006*: Generalized convulsive epilepsy, with intractable epilepsy, *0.005*: Epilepsy, unspecified, with intractable epilepsy, *0.004*: Infantile spasms, without mention of intractable epilepsy, ...
0.131: Other convulsions	**(2)** *0.083*: Other convulsions, *0.015*: Convulsions
0.055: Downs syndrome	**(1)** *0.001*: Conditions due to anomaly of unspecified chromosome
0.046: Intellectual disability	**(1)** *0.034*: Intellectual disability
0.040: Other Disorders of the Central Nervous System	**(31)**, including: *0.052*: Epilepsy, unspecified, without mention of intractable epilepsy, *0.006*: Generalized convulsive epilepsy, with intractable epilepsy, *0.002*: Other brain condition, *0.002*: Quadriplegia, *0.0001*: Hemiplegia, unspecified, affecting dominant side, *0.0001*: Migraine without aura, with intractable migraine, *0.00009*: Flaccid hemiplegia Flaccid hemiplegia and hemiparesis affecting unspecified side, *0.00005*: Metabolic encephalopathy...

Subject Headings (MeSH) (Lipscomb 2000). These terms are hierarchical: terms near the root are more general than those further down the tree. For example, *cardiovascular diseases* subsumes *heart diseases*, which is in turn a parent of *Heart Aneurysm*.

MeSH terms are commonly used in *systematic reviews* (SR) (Grimshaw and Russell 1993), where the goal is to summarize all publications relating to precise, scientific questions. Retrieving these articles is time-consuming, expensive, and tedious. MeSH terms can help researchers quickly decide if an article is relevant. MeSH terms are manually assigned to articles by an expert team of annotators, which results in high variability with respect to the specificity of the terms. GS-LDA provides a means of identifying latent concepts that define distributions over terms nearby in the MeSH structure. These interpretable, sparse topics can provide concise summaries of biomedical documents, easing the evidence retrieval process for overburdened researchers and physicians.

We consider a dataset of 1218 documents annotated with 5347 unique MeSH terms (23 average terms per document) that were screened for a systematic review of the effects of calcium-channel blocker (CCB) drugs (Cohen et al. 2006). Figure 3e shows that the test log-likelihood for GS-LDA on these data is on par with LIDA, but the model produces a much sparser summary of concept-words (figure 3f). Here, the concepts found by GS-LDA correspond to sets of MeSH terms that might help researchers rapidly identify studies reporting results for trials investigating the use of CCB's – without having to make sense of a topic comprising hundreds of unique MeSH terms.

Table 2 shows the top concept-words in a sample topic discovered by GS-LDA compared to a similar topic discovered by LIDA. GS-LDA gives most of topic mass to double-blind trials and CCBs; knowing the relative prevalence in an article of this topic would clearly help a researcher looking to find reports of randomized controlled trials of CCBs. In contrast, words related to concept CCBs are divided among terms in LIDA. Some of the LIDA terms, such as *Drug Therapy, Combination* and *Mibefradil* are also present in GS-LDA, but with much lower probability – the concept CCB summarizes most of the instances. We note that a professional systematic reviewer at [Anonymous] confirmed that the more concise topics found by GS-LDA would be more useful in facilitating evidence retrieval tasks than those found by LIDA.

Related Work

GS-LDA is a novel approach to inducing interpretability by exploiting structured vocabularies. Prior work on interpretable topic models has focused on various notions of coherence. (Chang et al. 2009) measured interpretability by how easily a human could identify an inserted "intruder" word among the top 5 words in a topic. The ease of intrusion detection was *negatively* correlated with test likelihood. (Newman et al. 2010; Mimno et al. 2011) introduced measures of coherence that correlate with human annotations of topic quality. The evaluations in all of these works focussed only on the top-n word lists, a powerful indication of closely linked sparsity is to interpretability.

In contrast, our approach does not sacrifice predictive quality and, by using the ontological structure, provides a compact summary that describes *most* of the words, not just the top n. This quality is particularly valuable when annotation disagreement or diagnostic "slosh" can result in a large number of words with non-trivial probabilities. The use of a human-provided structure to induce interpretability distinguishes GS-LDA from other hierarchical topic models where the structure is learned (Blei, Griffiths, and Jordan 2010; Chen, Dunson, and Carin 2011; Li and McCallum 2006). Most similar to our work is the *super-word* concept modeling of (El-Arini, Fox, and Guestrin 2012), in which

Table 2: A sample from the Calcium Channels systematic review. Superscripts denote the same term found at different levels in the MeSH structure; we collapse them when they appear sequentially in a topic. GS-LDA captures the concepts "double-blind trial" and "calcium channel blockers" in one topic, which is exactly what the researchers were looking to summarize in this systematic review.

Graph-Sparse LDA (21 total nonzero)	LIDA (90 total nonzero)
0.565: Double-Blind Method^{1,2,3,4}	(**1**) *0.353*: Double-Blind Method^{1,2,3,4}
0.110: Calcium Channel Blockers^{1,2}	(**7**) *0.031* Adrenergic beta-Antagonists^{1}, *0.026* Drug Therapy, Combination, *0.022* Calcium Channel Blockers, *0.016* Felodipine, *0.015* Atenololm, *0.006* Benzazepines, *0.01* Mibefradil^{1,2}
0.095: Angina Pectoris^{1,2}	(**3**) *0.030*: Angina Pectoris^{2}, *0.030*: Myocardial Ischemia^{1,2}, *0.003*: Atrial Flutter

auxiliary information about the words, encoded in a feature vector, can be used to encourage or discourage words from being part of the same concept. Unlike (El-Arini, Fox, and Guestrin 2012), we use the graph structure to guide the formation of concepts, which maintains interpretability without requiring concepts to have sparse support. More generally, while a learned hierarchical structure allows for statistical sharing between topics, each topic is still a distribution over a large vocabulary. The interpretation task is more complex as the expert must now inspect both the hierarchy and the topics.

Finally, curated ontologies have been used in other topic modeling contexts. (Abney and Light 1999) uses hierarchies for word-sense disambiguation in n-gram tuples, and (Boyd-Graber, Blei, and Zhu 2007) incorporate this idea into topic models. (Slutsky, Hu, and An 2013; Andrzejewski, Zhu, and Craven 2009) use the hierarchical structure as partial supervision in topic models to improve predictive performance. In contrast to these efforts, which focus on prediction, Graph-Sparse LDA uses the ontology in a probabilistic—rather than enforced—manner to obtain sparse, interpretable topics.

Discussion and Conclusion

Topic models have revolutionized many prediction tasks, and scientists now commonly use them to uncover understandable structure from data. *Understanding* is a more nuanced objective than prediction, and successful applications of topic models for this aim cannot ignore the structured knowledge-bases that exist in many scientific domains.

Graph-Sparse LDA exploits such resources to induce interpretable topics. Specifically, leveraging ontological knowledge enabled us to uncover sparse sets of concept words that provided succinct, interpretable topic summaries while still explaining a large number of observed words. Our approach is robust in the sense that if ontology encoded in the concept-word matrix P does not permit sparse solutions, then we will simply discover a less sparse topic-concept matrix A that is still predictive. The combination of this representational power and a novel, efficient inference procedure allowed us to realize topic interpretability while matching (or exceeding) state-of-the-art predictive performance.

While we have focused on biomedical domains, our approach could be applied to general text corpora using standard hierarchies such as WordNet. Our model can consider concept-words that generate any nearby observed word, where the definition of "nearby" (i.e., the sparsity of concept-word matrix P) is entirely up to the model designer. Thus, the underlying structure can be a tree, a DAG, or just some collection of neighborhoods. The key benefit of using our approach is that the model designer can now easily view and distinguish between intra-concept correlations (in the concept-word matrix P, presumably unsurprising) and inter-concept correlations (in the topic-concept matrix A, potentially new information). Finally, our inference procedure is simpler than many other hierarchical models. We expect Graph-Sparse LDA to be useful for a variety of topic discovery applications in which the observed dimensions have human-understandable relationships.

Acknowledgments

We are grateful for Isaac Kohane for his many discussions and the i2b2 team at Boston Children's Hospital for providing us with the ASD data.

References

Abney, S., and Light, M. 1999. Hiding a semantic hierarchy in a Markov model. In *Workshop on Unsupervised Learning in Natural Language Processing*, 1–8.

Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S. M.; and Telgarsky, M. 2012. Tensor decompositions for learning latent variable models. http://arxiv.org/abs/1210.7559.

Andrzejewski, D.; Zhu, X.; and Craven, M. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *ICML*, 25–32.

Archambeau, C.; Lakshminarayanan, B.; and Bouchard, G. 2011. Latent IBP compound Dirichlet allocation. In *NIPS Bayesian Nonparametrics Workshop*.

Blei, D. M.; Griffiths, T. L.; and Jordan, M. I. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM* 57(2):7:1–7:30.

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

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32:D267–D270.

Boyd-Graber, J. L.; Blei, D. M.; and Zhu, X. 2007. A topic model for word sense disambiguation. In *EMNLP-CoNLL*, 1024–1033.

Chang, J.; Boyd-Graber, J. L.; Gerrish, S.; Wang, C.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*, 288–296.

Chen, H.; Dunson, D. B.; and Carin, L. 2011. Topic modeling with nonparametric Markov tree. In *ICML*, 377–384.

Cohen, A. M.; Hersh, W. R.; Peterson, K.; and Yen, P.-Y. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13(2):206–219.

Doshi-Velez, F.; Ge, Y.; and Kohane, I. 2013. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*.

El-Arini, K.; Fox, E. B.; and Guestrin, C. 2012. Concept modeling with superwords. http://arxiv.org/abs/1204.2523.

Fei-Fei, L., and Perona, P. 2005. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, 524–531.

Ghassemi, M.; Naumann, T.; Joshi, R.; and Rumshisky, A. 2012. Topic models for mortality modeling in intensive care units. In *ICML 2012 Machine Learning for Clinical Data Analysis Workshop*.

Griffiths, T., and Ghahramani, Z. 2011. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research* 12:1185–1224.

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

Grimshaw, J. M., and Russell, I. T. 1993. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *The Lancet* 342(8883):1317–1322.

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

Lipscomb, C. E. 2000. Medical subject headings (MeSH). *Bull Med Libr Assoc.* 88(3): 265266.

Mimno, D.; Wallach, H.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *EMNLP*.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, 100–108. Association for Computational Linguistics.

Slutsky, A.; Hu, X.; and An, Y. 2013. Tree labeled LDA: A hierarchical model for web summaries. In *IEEE International Conference on Big Data*, 134–140.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. *Handbook of latent semantic analysis* 427(7):424–440.

Wang, C., and Blei, D. 2009. Decoupling Sparsity and Smoothness in the Discrete Hierarchical Dirichlet Process. In Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, C. K. I.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 22*. 1982–1989.

Wang, Y., and Mori, G. 2009. Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(10):1762–1774.

Williamson, S.; Wang, C.; Heller, K. A.; and Blei, D. M. 2010. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 1151–1158.