Source-LDA: Enhancing probabilistic topic models using prior knowledge sources

Justin Wood[1,2], Patrick Tan[1], Ariyam Das[1], Wei Wang[1], Corey Arnold[2]

[1]Department of Computer Science, University of California Los Angeles, CA 90095, USA

[2]Medical Imaging Informatics Group, University of California, Los Angeles, CA 90095, USA

{juwood03, patrickptt, cwarnold}@ucla.edu, {ariyam, weiwang}@cs.ucla.edu

Abstract—**Topic modeling has increasingly attracted interests from researchers. Common methods of topic modeling usually produce a collection of unlabeled topics where each topic is depicted by a distribution of words. Associating semantic meaning with these word distributions is not always straightforward. Traditionally, this task is left to human interpretation. Manually labeling the topics is unfortunately not always easy, as topics generated by unsupervised learning methods do not necessarily align well with our prior knowledge in the subject domains. Two approaches have been taken to address this issue, either by a post-processing procedure that assigns each topic with a label from a prior knowledge base which is semantically closest to the word distribution of the topic, or by a supervised topic modeling approach that restricts the topics to a predefined set whose word distributions are provided a priori. Neither approach is ideal in that the former one may produce labels that do not accurately describe the word distributions, and the latter one lacks the ability to detect unknown topics that are crucial to enrich our knowledge base. Our goal in this paper is to introduce a semi-supervised Latent Dirichlet allocation (LDA) model, Source-LDA, which incorporates prior knowledge to guide the topic modeling process to improve the quality of resulting topics, as well as their labeling. This is done by integrating existing labeled knowledge sources representing known potential topics into a probabilistic topic model. These knowledge sources are translated into a distribution and used to set the hyperparameters of the Dirichlet generated distribution over words. This approach ensures that the topic inference process is consistent with existing knowledge, and simultaneously, allows for discovery of new topics. The results show both improved topic generation and accurate topic labeling superior to those obtained using various labeling approaches based off LDA.**

I. INTRODUCTION

Existing topic modeling is often based off of Latent Dirichlet allocation (LDA) [1] and involves analyzing a given corpus to produce a distribution over words for each latent topic and a distribution over latent topics for each document. Although the distributions representing topics have been shown to be useful and generally representative of a linguistic topic, assigning labels to topics is often left to manual interpretation.

The identification of topic labels in the general sense is useful in summarizing a set of words by a single label. For example, words such as pencil, laptop, ruler, eraser, and book can be mapped to the label "School Supplies." Adding descriptive semantics to each topic can help people, especially those without domain knowledge, to understand topics obtained by topic modeling.

A motivating application of accurate topic labeling is that of summarization systems for primary care physicians, who are faced with the challenges of being inundated with too much

data for a patient and too little time to comprehend it all [2]. The labels can be used to more appropriately and quickly identify a patient's medical history, leading to better outcomes for the patient. This added information can bring significant value to the field of clinical informatics which already utilizes topic modeling without labeling [3]–[5], as well as many other domains.

Existing approaches in labeling topics usually do their fitting of labels to topics after completion of the unsupervised topic modeling process. A topic produced by this approach may not always match well with any semantic concepts and is therefore hard to be categorized by a single label. These problems are best illustrated via a simple case study.

1) Case Study: Suppose a corpus of a news source that consists of two articles is given by documents $d1$ and $d2$ each with three words:

d1 - pencil, pencil, umpire

d2 - ruler, ruler, baseball

LDA (with the traditionally used collapsed Gibbs sampler, standard hyperparameters and the number of topics (K) set as two) would output different results for different runs due to the inherent stochastic nature. It is very possible to obtain the following result of topic assignments:

d1 - pencil[1], pencil[1], umpire[2]

d2 - ruler[2], ruler[2], baseball[1]

But these assignments to topics differs from the ideal solution that involves knowing the context of the topics in which these words come from. If the topic modeling was to incorporate prior knowledge about the topics "School Supplies" and "Baseball" then a topic modeling process can more likely generate the ideal topic assignments of:

d1 - pencil[2], pencil[2], umpire[1]

d2 - ruler[2], ruler[2], baseball[1]

and also assign a label of "School Supplies" to topic 1 and "Baseball" to topic 2.

To overcome this problem, one may take a supervised approach that incorporates such prior knowledge into the topic modeling process in order to improve the quality of topic assignments and effectively label topics. However, existing supervised approaches [6]–[8] are either too lenient or too

strict. In the Concept-topic model (CTM) [6], for example, a multinomial distribution is placed over known concepts with associated word sets. This pioneering approach does integrate prior knowledge, but does not take into account word distributions. For example if a document is generated about the topic "School Supplies" it is much more probable to see the word "pencil" than the word "compass" even though both words may be associated with the topic "School Supplies". This technique also requires some supervision which requires manually inputting preexisting concepts and their bags of words.

Another approach given by Hansen et al. as explicit Dirichlet allocation [7] incorporates a preexisting distribution based off of Wikipedia but does not allow for variance from the Wikipedia distribution. This approach fulfills the goal of incorporating prior knowledge with their distributions but requires the topic in the generated corpus strictly follow the Wikipedia word distributions.

To address these limitations, we propose the Source-LDA model which is a balance between these two approaches. The goal is to allow for simultaneous discovery of both known and unknown topics. Given a collection of known topics and their word distributions, Source-LDA is able to identify the subset of these topics that appear in a given corpus. It allows some variance in word distributions to the extent that it optimizes the topic modeling. A summary of the contributions of this work are:

1) We propose a novel technique to topic modeling in a semi-supervised fashion that takes into account preexisting topic distributions.
2) We show how to find the appropriate topics in a corpus given an input set that contains a subset of the topics used to generate a corpus.
3) We explain how to make use of prior knowledge sources. In particular, we show how to use Wikipedia articles to form word distributions.
4) We introduce an approach that allows for variance from an input topic to the latent topic discovered during the topic modeling process.

The rest of this paper is organized as follows: In Section 2, we give a brief introduction to the LDA algorithm and the Dirichlet distribution. A more detailed description of the Source-LDA algorithm is presented in Section 3. In Section 4, the algorithm is used and evaluated under various metrics. Related literature will be highlighted in Section 5. Section 6 gives the conclusions of this paper.

For reproducible research, we make all of our code available online.[1]

II. PRELIMINARIES

A. Dirichlet Distribution

The Dirichlet distribution is a distribution over probability mass functions with a specific number of atoms and is commonly used in Bayesian models. A property of the Dirichlet that is often used in inference of Bayesian models is conjugacy to the multinomial distribution. This allows for the posterior of a random variable with a multinomial likelihood and a Dirichlet prior to also be a Dirichlet distribution.

The parameters are given as a vector denoted by α. The probability density function for a given probability mass function (PMF) θ and parameter vector α of length J is defined as:

$$f(\theta, \alpha) = \frac{\Gamma(\sum_i^J \alpha_i)}{\prod_i^J \Gamma(\alpha_i)} \prod_i^J \theta_i^{\alpha_i - 1}$$

A sample from the Dirichlet distribution produces a PMF that is parameterized by α. The choice of a particular set of α values influences the outcome of the generated PMF. If all α values are the same (symmetric parameter), as α approaches 0 the probability will be concentrated on a smaller set of atoms. As α approaches infinity then the PMF will become the uniform distribution. If all α_i are natural numbers then each individual α_i can be thought of as "virtual" count for the i_{th} value [9].

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is the basis for many existing probabilistic topic models, and the framework for the approach presented by this paper. Since we enhance the LDA model in our proposed approach it is worth giving a brief overview of the algorithm and model of LDA.

LDA is a hierarchical Bayes model which utilizes Dirichlet priors to estimate the intractable latent variables of the model. At a high level, LDA is based on a generative model in which each word of an input document from a corpus is chosen by first selecting a topic that corresponds to that word and then selecting the word from a topic-to-word distribution. Each topic-to-word distribution and word-to-topic distribution is drawn from its respective Dirichlet distribution. The formal definition of the generative algorithm over a corpus is:

1. For each of the K topics ϕ_k:
2.   Choose $\phi_k \sim \mathrm{Dir}(\beta)$
3. For each of the D documents d:
4.   Choose $N_d \sim \mathrm{Poisson}(\xi)$
5.   Choose $\theta_d \sim \mathrm{Dir}(\alpha)$
6.   For each of the N_d words $w_{n,d}$:
7.   Choose $z_{n,d} \sim \mathrm{Multinomial}(\theta)$
8.   Choose $w_{n,d} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$

From the generative algorithm the resultant Bayes model is shown by Figure 1(a).

Bayes' law is used to infer the latent θ distribution, ϕ distribution, and z

$$P(\theta, \phi, z | w, \alpha, \beta) = \frac{p(\theta, \phi, z, w | \alpha, \beta)}{p(w | \alpha, \beta)}$$

Unfortunately the computation of this equation exactly is intractable. Hence, it must be approximated with techniques such as expectation-maximization [1], Gibbs sampling or collapsed Gibbs sampling [10].

III. PROPOSED APPROACH

Source-LDA is an extension of the LDA generative model. In Source-LDA, after a known set of topics are identified, an initial word-to-topic distribution is generated from Wikipedia
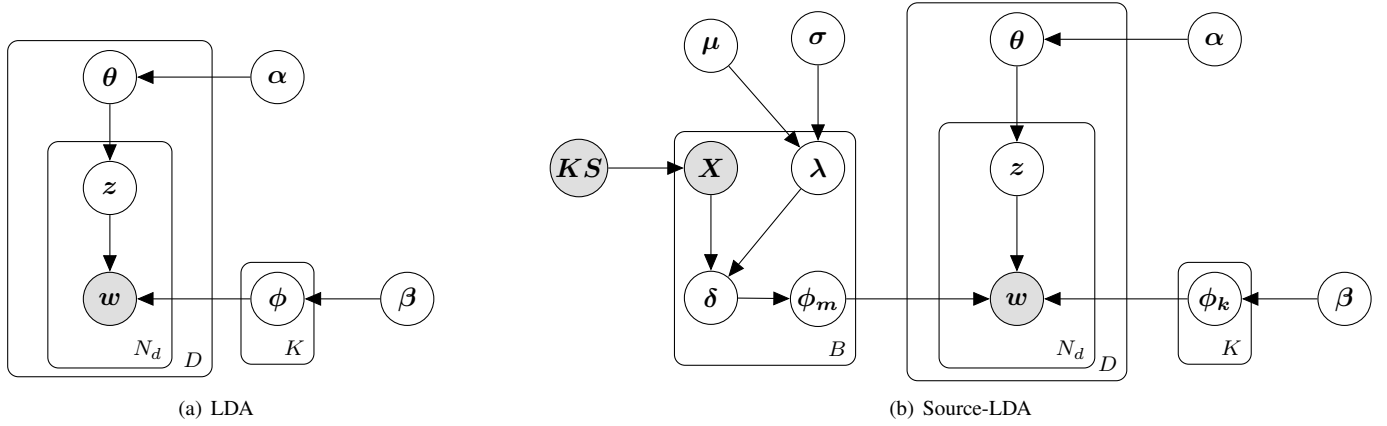

Fig. 1: Plate notation for LDA (a), and the proposed Source-LDA (b).

articles. The desiderata is to enhance existing LDA topic modeling by integrating prior knowledge into the topic modeling process. The relevant terms and concepts used in the following discussion are defined below.

Definition 1 (Knowledge source): A knowledge source is a collection of documents that are focused on describing a set of concepts. For example the knowledge source used in our experiments are Wikipedia articles that describe the categories we select from the Reuters dataset.

Definition 2 (Source Distribution): The source distribution is a discrete probability distribution over the words of a document describing a topic. The probability mass function is given by

$$\forall w_i \in W, \ f(w_i) = \frac{n_{w_i}}{\sum_j^G n_{w_j}}$$

where W is the set of all words in the document, $G = |W|$, and n_{w_i} is the number of times word w_i appears in the document.

Definition 3 (Source Hyperparameters): For a given document in a knowledge source the knowledge source hyperparameters are defined by the vector $(X_1, X_2, \ldots, X_V)$ where $X_i = n_{w_i} + \epsilon$ and ϵ is a very small positive number. V is the size of the vocabulary of the corpus for which we are topic modeling, and n_{w_i} is the number of times the word w_i from the corpus vocabulary appears in the knowledge source document.

Three approaches are given that capture the intent of Source-LDA, the first being a simple enhancement to the LDA model that allows for the influencing of topic distributions, but suffers from the limitation of needing more user intervention. The second approach allows for the mixing of unknown topics and the third approach combines the previous two approaches and moves toward a complete solution to topic modeling based off of prior knowledge sources.

A. Bijective Mapping

In the simplest approach, the Source-LDA model assumes that there exists a 1-to-1 mapping between a known set of topics and the topics used to generate a corpus. The generative model then assumes that, instead of selecting topic-to-word distributions from sampling from the Dirichlet distribution, a set of K distributions are given as input and sampled from

after each topic assignment is sampled for a given token position. The generative process for a corpus adapted from the traditional LDA generative model during the construction of the ϕ distributions is as follows (for brevity only the relevant parts of the existing LDA algorithm are shown):

1. For each of the K topics ϕ_k:
2. $\quad \delta_k \leftarrow (X_{k,1}, X_{k,2}, \ldots, X_{k,V})$
3. $\quad$ Choose $\phi_k \sim \text{Dir}(\delta_k)$

Where $(X_{k,1}, X_{k,2}, \ldots, X_{k,V})$ represents the knowledge source hyperparameters for the k_{th} knowledge source document. The generative model only differs from the traditional LDA model in how each ϕ is built. Therefore the derivation for inference is a simple factor as well. To approximate the distributions for θ and ϕ a collapsed Gibbs sampler can approximate the z assignments as follows:

$$P(w_i|z_i{=}j, z_{\text{-}i}, w_i) \propto P(w_i|z_i{=}j, z_{\text{-}i}, w_{\text{-}i})P(z_i{=}j|z_{\text{-}i})$$

From the Bayesian Model the following equations can be easily be generated

$$P(w_i|z_i{=}j, z_{\text{-}i}, w_{\text{-}i}) = \int P(w_i|z_i{=}j, \phi_j)P(\phi_j|z_{\text{-}i}, w_{\text{-}i})d\phi_j$$

with

$$P(\phi_j|z_{\text{-}i}, w_{\text{-}i}) \propto P(w_i|\phi_j, z_{\text{-}i})P(\phi_j)$$

$$P(\phi_j|z_{\text{-}i}, w_{\text{-}i}) = Dir(\delta_{i,j} + n_{w_{\text{-}i,j}})$$

$$P(w_i|z_i{=}j, \phi_j) = \phi_{w_{i,j}}$$

$$P(w_i|z_i{=}j, z_{\text{-}i}, w_{\text{-}i}) = Dir(\delta_{i,j} + n_{w_{\text{-}i,j}}) \int \phi_{w_{i,j}} d\phi_j$$

$$P(w_i|z_i{=}j, z_{\text{-}i}, w_{\text{-}i}) = \frac{n_{\text{-}i,j}^{w_i} + \delta_{i,j}}{n_{\text{-}i,j}^{(\cdot)} + \sum_a^V \delta_{a,j}}$$

For brevity since the prior probability is unchanged in the "Bijective Mapping" model we will skip the derivation which is well defined in other articles [10]–[12].

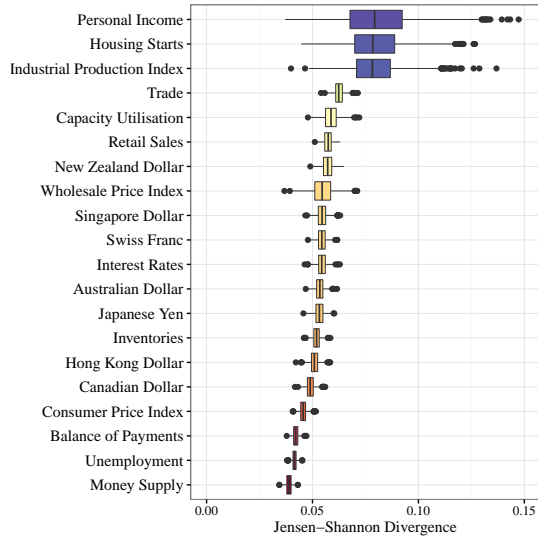$$P(z_i{=}j|z_{\text{-}i}) = \frac{n_{\text{-}i,j}^{d_i} + \alpha}{n_{\text{-}i}^{(d_i)} + K\alpha}$$

Fig. 2: Box plots [13] showing the Jensen-Shannon divergence (the JS divergence measures the distance or similarity between probablity distributions) of 1000 Dirichlet samples parameterized by source hyperparameters for a subset of knowledge source topics. The topics were taken from Wikipedia pages.

Putting the two equations together gives the final Gibbs sampling equation:

$$P(z_i{=}j|z_{-i},w) \propto \frac{n_{-i,j}^{w_i} + \delta_{i,j}}{n_{-i,j}^{(\cdot)} + \sum_a^V \delta_{a,j}} \frac{n_{-i,j}^{d_i} + \alpha}{n_{-i}^{(d_i)} + K\alpha}$$

Given the approximation to the topic assignments, the θ and ϕ distributions can be calculated as:

$$\phi_{w,t} = \frac{n_{w,t} + \delta_{w,t}}{n_t + \sum_a^V \delta_{a,t}}$$

$$\theta_{t,d} = \frac{n_{d,t} + \alpha}{n_d + K\alpha} \qquad (1)$$

In the case when all topics are known this model has the advantage of conforming the ϕ distributions to the source distributions, but has three drawbacks. First, even though there is some variability between the ϕ distribution and source distribution, as illustrated by Figure 2, there may be cases in which this constraint should be relaxed even further. This is because it is entirely possible to generate a corpus about a known topic without exactly following the frequencies at which the topic is discussed in its respective article. This model also requires the user to input the known topics, and other possible supervised approaches may be better suited to the task [14]–[16]. The third drawback is that we are not allowing the possibility that the corpus was generated from a mixture of known topics and unknown topics, which is a more realistic scenario for an arbitrary document. The next model aims to resolve this last deficiency.

B. Known Mixture of Topics

The next model assumes that in the topic model it is given how many topics are known topics (as well as their word distributions) and how many are unknown topics. The previous approach works quite well in this situation in that an unknown topic will have a symmetric beta parameter which will capture assignments which were unallocated due to a low probability in matching any known topic.

The resulting model helps to solve the existing problems of the bijective model and only requires a minor input to the existing generative model. The resulting model works quite well with the bijective model in that the symmetric Dirichlet prior can be used to guide a topic toward being a general unknown topic or a known topic. The model changes as shown below with a minor change to the generative algorithm and the collapsed Gibbs sampling.

1.   For each of the K topics ϕ_k:
2.   **if** $k \leq T$ **then**
3.   Choose $\phi_k \sim \text{Dir}(\beta)$
4.   **else**
5. $\delta_k \leftarrow (X_{k,1}, X_{k,2}, \ldots, X_{k,V})$
6.   Choose $\phi_k \sim \text{Dir}(\delta_k)$

Where T is the total number of non-source topics. The change required to the collapsed Gibbs sampling is then:

$$P(z_i{=}j|z_{-i},w) \propto \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{d_i} + \alpha}{n_{-i}^{(d_i)} + K\alpha}, \ \forall i \leq T \qquad (2)$$

and

$$P(z_i{=}j|z_{-i},w) \propto \frac{n_{-i,j}^{w_i} + \delta_{i,j}}{n_{-i,j}^{(\cdot)} + \sum_a^V \delta_{a,j}} \frac{n_{-i,j}^{d_i} + \alpha}{n_{-i}^{(d_i)} + K\alpha}, \ \forall i > T$$

This approach gives the benefit of allowing a mixture of known topics and unknown topics, but problems still arise in that the Dirichlet distributions for the source distribution may be too restricting.

C. Source-LDA

By using the counts as hyperparameters, the resultant ϕ distribution will take on the shape of the word distribution derived from the knowledge source. But this may be counter to the aim of enhancing existing topic modeling. With the goal to influence the ϕ distribution, it is entirely plausible for there to be divergence between the two distributions. In other words, ϕ may not need to strictly follow the corresponding knowledge source distribution.

1) Variance from the source distribution: To allow for this relaxation, another parameter λ is introduced into the model which is used to allow for a higher deviance from the source distribution. To obtain this variance each source hyperparameter will be raised to a power of λ. Thus as λ approaches 0 each hyperparameter will approach 1 and the subsequent Dirichlet draw will allow all discrete distributions with equal probability. As λ approaches 1 the Dirichlet draw will be tightly conformed to the source distribution.

The addition of λ changes the existing generative model only slightly and allows for a variance for each individual δ_i, which frees us from an overly restrictive binding to the associated knowledge source distribution. The λ parameter acts as a measure of how much divergence is allowed for a given modeled topic from the knowledge source distribution. Figure
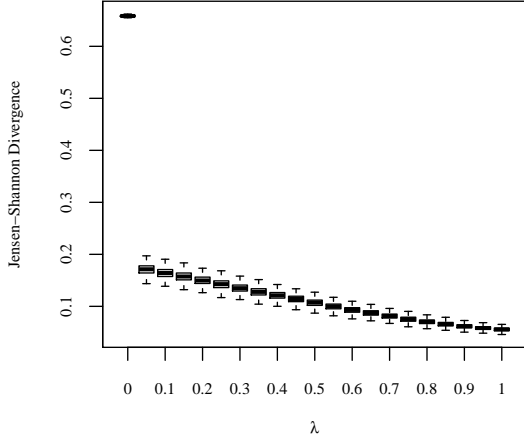
Fig. 3: Box plots showing how the JS divergence between a source distribution and a Dirichlet sample parameterized by source hyperparameters raised to λ changes with changes to λ without smoothing.
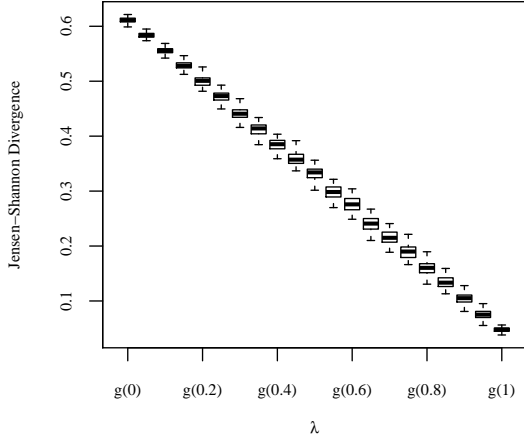

Fig. 4: The JS divergence between a source distribution and a Dirichlet sample parameterized by source hyperparameters raised to λ with λ mapped to a linear smoothing function g.

3 shows how the JS Divergence changes with changes to the λ parameter.

5. $$\delta_k \leftarrow [(X_{k,1})^\lambda, (X_{k,2})^\lambda, \ldots, (X_{k,V})^\lambda]$$

With the introduction of λ as an input parameter, the new topic model has the advantage of allowing variance and also leaves the collapsed Gibbs sampling equation unchanged. However this also requires a uniform variance from the knowledge base distribution for all latent topics. This can be a problem if the corpus was generated with some topics influenced strongly while others less so. To solve this we can introduce λ as a hidden parameter of the model.

2) Approximating λ: In the ideal situation λ will be as close to 1 for most knowledge based latent topics, with the flexibility to deviate as required by the data. For this we assume a Gaussian prior over λ with mean set to μ. The variance then becomes a modeled parameter that conceptually can be thought of as how much variance from the knowledge source distribution we wish to allow in our topic model. In assuming a

Gaussian prior for λ, we must integrate λ out of the collapsed Gibbs sampling equations (only the probability of w_i under topic j is shown, the probability of topic j in document d is unchanged and omitted).

$$P(z_i{=}j|z_{-i}, w) \propto \int \frac{n_{-i,j}^{w_i} + (\delta_{i,j})^\lambda}{n_{-i,j}^{(\cdot)} + \sum_a^V (\delta_{a,j})^\lambda} \mathcal{N}(\mu, \sigma) d\lambda$$

ϕ then becomes

$$\phi_{w,t} = \int \frac{n_{w,t} + (\delta_{w,t})^\lambda}{n_t + \sum_a^V (\delta_{a,t})^\lambda} \mathcal{N}(\mu, \sigma) d\lambda$$

Unfortunately closed form expressions for these integrals are hard to obtain and so they must be approximated numerically during sampling.

Another problem arises in that the change of λ is not in par with the change of the Gaussian distribution, as can be seen in Figure 3. To make the changes of λ more in line with that expected from the Gaussian PDF, we must map each individual λ value in the range 0 to 1 with a value which produces a change in the JS divergence in a linear fashion. We approximate a function, $g(x)$ with a linear derivative, shown in Figure 4. The approach taken to approximate $g(x)$ is by linear interpolation of an aggregated large number of samples for each point taken in the range 0 to 1. Our collapsed Gibbs sampling equations then becomes:

$$P(z_i{=}j|z_{-i}, w) \propto \int \frac{n_{-i,j}^{w_i} + (\delta_{i,j})^{g(\lambda)}}{n_{-i,j}^{(\cdot)} + \sum_a^V (\delta_{a,j})^{g(\lambda)}} \mathcal{N}(\mu, \sigma) d\lambda \quad (3)$$

$$\phi_{w,t} = \frac{n_{w,t} + \beta}{n_t + V\beta}, \ \forall t \leq T \quad (4)$$

and

$$\phi_{w,t} = \int \frac{n_{w,t} + (\delta_{w,t})^{g(\lambda)}}{n_t + \sum_a^V (\delta_{a,t})^{g(\lambda)}} \mathcal{N}(\mu, \sigma) d\lambda, \ \forall t > T$$

3) Superset Topic Reduction: A third problem involves knowing the right mixture of known topics and unknown topics. It is also entirely possible that many known topics may not be used by the generative model. Our desire to leave the model as unsupervised as possible calls for input that is a superset of the actual generative topic selection in order to avoid manual topic selection. In the case of modeling only a specific number of topics over the corpus, the problem then becomes how to choose which knowledge source latent topics to allow in the model vs. how many unlabeled topics to allow.

The goal then is to allow for a superset of knowledge source topics as input and then during the inference to select the best subset of these with a mixture of unknown topics where the total number of unlabeled topics is given as input K. The approach given is to use a mixture of K unlabeled topics alongside the labeled knowledge source topics. The total number of topics then becomes T. During the inference we eliminate topics which are not assigned to any documents. At the end of the sampling phase we then can use a clustering algorithm (such as k-means, JS divergence) to further reduce the modeled topics and give a total of K topics. As described more in the experimental section, with the goal of capturing topics that were frequently occurring in the corpus, topics not appearing in a frequent enough of documents were eliminated.

The complete generative process is shown in Figure 1(b)

and described below:

1. For each of the T topics ϕ_t:
2.   **if** $t \leq K$ **then**
3.   Choose $\phi_t \sim \text{Dir}(\beta)$
4.   **else**
5.   Choose $\lambda_t \sim \mathcal{N}(\mu, \sigma)$
6. $\delta_t \leftarrow [(X_{t,1})^{g(\lambda_t)}, (X_{t,2})^{g(\lambda_t)}, \ldots, (X_{t,V})^{g(\lambda_t)}]$
7.   Choose $\phi_t \sim \text{Dir}(\delta_t)$
8. For each of the D documents d:
9.   Choose $N_d \sim \text{Poisson}(\xi)$
10.   Choose $\theta_d \sim \text{Dir}(\alpha)$
11.   For each of the N_d words $w_{n,d}$:
12.   Choose $z_{n,d} \sim \text{Multinomial}(\theta)$
13.   Choose $w_{n,d} \sim \text{Multinomial}(\phi_{z_{n,d}})$

The full collapsed Gibbs sampling algorithm is given in algorithm 1.

4) Analysis: By using a clustering algorithm or thresholding the topic document frequency the collapsed Gibbs algorithm is guaranteed to produce K topics. The running time is a function of the number of iterations I, average words per document D_{avg}, number of documents D, number of topics T and number of approximation steps A, and is $\mathcal{O}(I \times D_{\text{avg}} \times D \times T \times A)$. This differs only from the traditional collapsed Gibbs sampling in LDA by an increase of $(T-K)A$. But since we have built the approach to potentially have a large $T - K$ this difference can have a significant impact on running times.

Approaches exist that can parallelize the sampling procedure, but these are often approximations or can potentially have slower than baseline running times [17]–[19]. We present two modifications to the original algorithm that allow for inference while guaranteeing the exactness of the results to the original Gibbs sampling. The first one makes use of prefix sums rules [20] and guarantees a running time of:

$$\mathcal{O}(I \times D_{\text{avg}} \times D \times A \times Max[T/P, P])$$

with P being the number of parallel units. This algorithm is given by Algorithm 2.

This algorithm is practical in situations where $T - K$ is large, but suffers from the limitations of the number of context switches required for the threads to wait at their respective barriers. A simpler implementation approach that reduces the number of context switches is to add the sums for each thread then wait for a barrier. When the barrier is released we add the end values together and then in parallel we add the remaining necessary items. This approach is given in Algorithm 3. The running time is then:

$$\mathcal{O}(I \times D_{\text{avg}} \times D \times A \times Max[T/P, P])$$

These two algorithms allow for mitigation of the increase in the number of topics and should approach times very similar to those of standard LDA runs. They are also very extensible and can be used in other optimization algorithms.

5) Input determination: Determining the necessary parameters and inputs into LDA is an established research area [21], but since the proposed model introduces additional input

---

Algorithm 1 Collapsed Gibbs

Input: Dirichlet hyperparameters α, β, a corpus C, vocabulary V, unlabeled topic count K, total topic count T, a set of source topics S, mean μ, variance σ, and iteration count I.
Output: θ, ϕ

procedure COLLAPSED_GIBBS(α, β, C, V, T)
    for $t = K + 1$ to T **do**
        Calculate g_t
    end for
    Initialize C_{topics} to random topic assignments
    Update n^w and n^d from C_{topics}
    for $iter = 1$ to I **do**
        for $i = 1$ to C **do**
            for $j = 1$ to $|C_i|$ **do**
                $C_{\text{topics}_{i,j}} \leftarrow Sample(i, j)$
            end for
        end for
    end for
    Calculate θ according to **Equation 1**
    Calculate ϕ according to **Equation 4**
    return θ, ϕ
end procedure

procedure SAMPLE(i, j)
    Decrement n^w and n^d accordingly
    for $t = 1$ to K **do**
        Calculate p_t according to **Equation 2**
    end for
    for $t = K + 1$ to T **do**
        Calculate p_t according to **Equation 3**
    end for
    $topic \sim \text{Multinomial}(p)$
    Increment n^w and n^d accordingly
    return $topic$
end procedure

Algorithm 2 Prefix Sums Parallel Sampling

procedure SAMPLE(i, j)
    Decrement n^w and n^d accordingly
    for i from 0 to $T - 1$ in parallel **do**
        if $i \leq K$ **then**
            Calculate p_i according to **Equation 2**
        else
            Calculate p_i according to **Equation 3**
        end if
        $p_i \leftarrow p_{i-1} + p_i$
    end for
    for d from 0 to $(\ln T) - 1$ **do**
        for i from 0 to $T - 1$ by 2^{d+1} in parallel **do**
            $p_{(i+2^{d+1}-1)} \leftarrow p_{(i+2^d-1)} + p_{(i+2^{d+1}-1)}$
        end for
    end for
    $p_{(T-1)} \leftarrow 0$
    for d from $(\ln T) - 1$ down to 0 **do**
        for i from 0 to $T - 1$ by 2^{d+1} in parallel **do**
            $h \leftarrow p_{(i+2^d-1)}$
            $p_{(i+2^{d+1}-1)} \leftarrow p_{(i+2^{d+1}-1)}$
            $p_{(i+2^{d+1}-1)} \leftarrow h + p_{(i+2^{d+1}-1)}$
        end for
    end for
    $topic \leftarrow$ Binary_Search(p)
    Increment n^w and n^d accordingly
    return $topic$
end procedure

---

requirements a brief overview will be given about how to best set the parameters and determine the knowledge source.

a) Parameter selection: To determine the appropriate parameters, techniques utilizing log likelihood have previously been established [10]. Since these approaches generally require held out data and are a function of the ϕ, θ, and α variables the introduction of λ and σ will not differentiate from their original equations. For example the perplexity calculations used for Source-LDA are based off of importance sampling [22], or latent variable estimation via Gibbs sampling [23]. Importance

Algorithm 3 Simple Parallel Sampling

```
procedure SAMPLE(i, j)
    Decrement n^w and n^d accordingly
    for i from 0 to T − 1 in parallel do
        if i ≤ K then
            Calculate p_i according to Equation 2
        else
            Calculate p_i according to Equation 3
        end if
        p_i ← p_{i−1} + p_i
    end for
    for i from 0 to T − 1 by T/P do
        p_i ← p_{(i−T/P)} + p_i
        ends_i ← p_i
    end for
    for i from 0 to T − 1 in parallel do
        diff ← p_end − ends_i
        p_i ← diff + p_i
    end for
    topic ← Binary_Search(p)
    Increment n^w and n^d accordingly
    return topic
end procedure
```

sampling is only a function of ϕ given by Equation 4, and estimation via Gibbs sampling can made using Equation 4 and by the following equation ($\tilde{z}$, $\tilde{w}$, and $\tilde{n}$ represent the corresponding variables in the test document set):

$$P(\tilde{z}_i{=}j|\tilde{z}_{-i}, \tilde{w}) \propto \frac{n_j^{w_i} + \tilde{n}_{-i,j}^{w_i} + \beta}{n_j^{(\cdot)} + \tilde{n}_{-i,j}^{(\cdot)} + W\beta} \frac{\tilde{n}_{-i,j}^{d_i} + \alpha}{\tilde{n}_{-i}^{(d_i)} + K\alpha}, \; \forall i \leq T$$

and

$$P(\tilde{z}_i{=}j|\tilde{z}_{-i}, \tilde{w}) \propto \frac{n_j^{w_i} + \tilde{n}_{-i,j}^{w_i} + \delta_{i,j}}{n_j^{(\cdot)} + \tilde{n}_{-i,j}^{(\cdot)} + \sum_a^V \delta_{a,j}} \frac{\tilde{n}_{-i,j}^{d_i} + \alpha}{\tilde{n}_{-i}^{(d_i)} + K\alpha}, \forall i > T$$

It is recommended to set the parameters so as to maximize the log likelihood. Further analysis such as whether or not the parameters can be learned a priori from the data are not the focus of this paper and are thus left as an open research area.

b) Knowledge source selection: Source-LDA is designed to be used only in with a corpus which has a known super set of topics which comprise a large portion of the tokens. An example of such a case is that of a corpus consisting of clinical patient notes. Since there are extensive knowledge sources comprising essentially all medical topics, Source-LDA can be useful in discovering and labeling these existing topics. In cases where it is not so easy to collect a superset of topics traditional approaches may be more useful.

IV. EVALUATION

To test the results of the Source-LDA algorithm we set up experiments to test against competing models. The most similar models to our proposed approach were used in comparison. These are: latent Dirichlet allocation (LDA) [1], explicit Dirichlet allocation (EDA) [7], and the Concept-topic model (CTM) [6]. Other approaches such as supervised latent Dirichlet allocation (sLDA) [14], discriminative LDA (DiscLDA) [15], and labeled LDA (L-LDA) [16] are not used since a main desiderata of Source-LDA is to require much less supervision than what is needed by these methods. Likewise hierarchical methods [24] are omitted because there is no established hierarchy in the knowledge source data for this model. We describe in more detail below the experimental setups and metrics used to compare results.
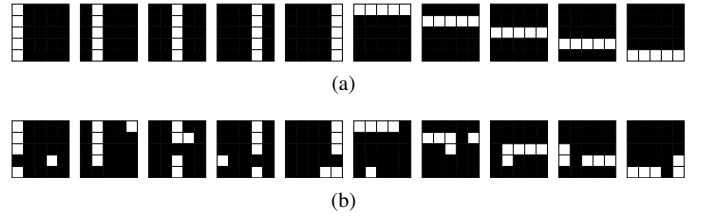

(a)

(b)

Fig. 5: A graphical representation of topics containing 1 word for the cell locations of row and column vectors in a 5 x 5 picture (a) and their augmented topics after swapping a random assigned word (pixel) with a random topic's assigned word (b).
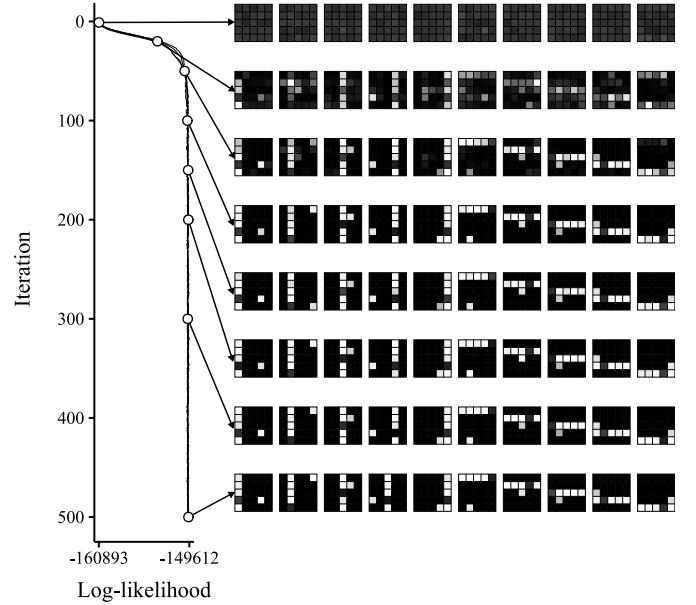

Fig. 6: Results from running Source-LDA for a corpus generated from topics in Figure 5(b) using a knowledge source of topics corresponding to Figure 5(a). Four separate runs are plotted to show the similarity of the log-likelihood relation to the iteration between the runs. The topics are shown visually at iteration 1, 20, 50, 100, 150, 200, 300 and 500 for a single run.

A. A Graphical Example

Following a previously established experiment [10], we show the utility of Source-LDA by visualizing topics created with words that correspond to the pixel locations in a 5×5 picture; but we add a key difference. The original topics are augmented, used to generate a corpus, and then hidden. Only the non augmented topics are given as input with the goal of discovering the augmented topics using the corpus and their original topics.

1) Experimental Setup: We start by creating ten topics with the vocabulary being the set of pixel locations in a 5×5 picture. The vocabulary (V) and bag of words representation of a topic (T_i) are defined as:

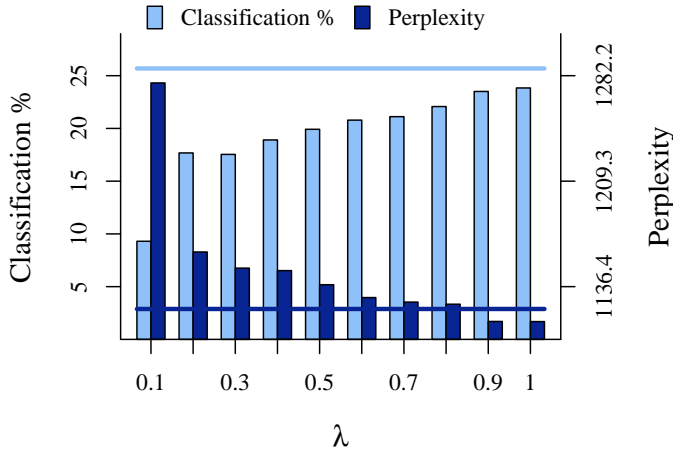$$V = \{xy \mid 0 \leq x < 5 \land 0 \leq y < 5\}$$

Fig. 7: Classification accuracy and perplexity values for fixed values of λ compared against the baseline values generated from a dynamic λ with a normal prior. The baseline values shown as lines represent the classification percentage of 25.7 and perplexity value of 1119.9

$$T_i = \begin{cases} xy \mid y = i \wedge 0 \le x < 5, & \text{if } 0 \le i < 5 \\ yx \mid y = i \wedge 0 \le x < 5, & \text{otherwise} \end{cases}$$

The topics are shown by Figure 5(a) with the intensity (I) of a pixel corresponding to word w in topic t equal to:

$$I(w,t) = Max[5 \times P(w|t), 1]$$

The representation of topics in this manner leads to a total of 10 topics. These original topics are then augmented by pairing each topic with a random different topic and swapping a random word (pixel) that is assigned to each topic given that the swapped words do not belong to their original assignments. Figure 5(b) shows the augmented topics which represent a 20% augmentation rate between the original topics. From the set of augmented topics we generate a 2,000 document corpus using the generative model of LDA. Each document consists of 25 words with topic assignments drawn from a distribution sampled from the Dirichlet distribution parameterized by $\alpha = 1$. With the knowledge source consisting solely of the original non augmented topics we run Source-LDA on the corpus hoping to discover and properly label the augmented topics.

2) Experimental Results: As shown in Figure 6, Source-LDA discovers the augmented topics given the set of original topics. Not only is Source-LDA able to find the topics correctly to the augmented distributions used in the generation of the corpus, but it is also able to match them to their respective non augmented source distributions. This simple experiment highlights a big advantage of Source-LDA; which is the ability to discover topics that differ from respective supervised input set. Other models such as EDA and CTM would be unable to label the augmented topics correctly due to the topics containing a word (pixel) not in the original distribution.

B. Integrating λ

A reasonable assumption to a corpus in which some topics are generated from a knowledge source is that the topics used in the corpus are going to deviate (more or less similar) from their respect source distributions and that each individual topic is going to deviate at a different rate than other topics.

The introduction of λ to Source-LDA as a parameter to be learned by the data allows the flexibility of different topics to be influenced differently by λ, but comes at an increase in computation cost. To show that in certain cases this flexibility is needed to obtain more accurate results we derive an experiment consisting of topics with different deviations from their respective source distributions.

1) Experimental Setup: A synthetic 500 document corpus is generated from a knowledge source of 100 randomly selected Wikipedia topics. The corpus is generated using the bijective model of Source-LDA as outlined in Section 3(A), consisting of 100 topics, an average word count per document of 100 words, $\mu = 0.5$, $\sigma = 1.0$ and $\alpha = 0.5$. Furthermore even though for each topic λ was drawn from $\mathcal{N}(\mu, \sigma^2)$ we bound the value drawn to the interval $[0, 1]$ for comparative analysis. We then run Source-LDA under the bijective model for a baseline of $\mu = 0.5$, $\sigma = 1.0$ against 10 runs of Source-LDA with λ fixed. After each run we compare the classification accuracy and perplexity values.

2) Experimental Results: For all fixed λ runs the baseline approach of varying λ in accordance with the normal distribution results in a higher classification accuracy. By allowing λ to deviate, the model can make up for incorrect parameter assignments due to a misleading perplexity value. As shown in Figure 7, classification accuracy is not perfectly correlated with perplexity. This is shown by the baseline method reporting a higher perplexity value than the fixed $\lambda = 1$ value while maintaining a higher classification accuracy. Even though we still recommend perplexity or other log-likelihood maximization approaches to set the parameters in any unknown data set, maximizing log-likelihood has been shown to be a less than perfect metric for evaluating topic models [25], [26]. In this experiment and the remaining experiments we take classification accuracy to be a more appropriate measurement for evaluating topic models.

C. Reuters Newswire Analysis

To show the type of topics discovered from Source-LDA we run the model on an existing dataset. This collection contains documents from the Reuters newswire from 1987. The dataset contains 21,578 articles, among a large set of categories. One important feature of the dataset are a set of given categories that we can use for our topic labeling. These include broad categories such as shipping, interest rates, and trade, as well as more refined categories such as rubber, zinc, and coffee. Our choice to apply our topic labeling method to this dataset is due to the fact that the Reuters dataset is widely used for information retrieval and text categorization applications. Due to its widespread use, it can considerably aid us in comparing our results to other studies. Additionally, because it contains distinct categories that we can use as our known set of topics, we can easily demonstrate the viability of our model.

1) Experimental Setup: Source-LDA, LDA, and CTM were run against the Reuters-21578 newswire collection. Since EDA does not discover new topics, nor does it update the word distributions of the input topics, we do not include EDA in this experiment. From the original 21,578 document corpus we select a subset of 2,000 documents. The Source-LDA and CTM supplementary distributions were generated by first obtaining a list of topics from the Reuters-21578 dataset. Next, for each topic, the corresponding Wikipedia article was

Retail Sales			Natural Gas			Japanese Yen		
SRC-LDA	IR-LDA	CTM	SRC-LDA	IR-LDA	CTM	SRC-LDA	IR-LDA	CTM
sales	said	loss	gas	corp	added	yen	japan	billion
mln	department	profit	natural	contract	plans	japan	u.s.	stabilisation
dlrs	u.s.	pretax	used	company	new	dollar	japanese	cds
year	parts	scale	water	services	copper	japanese	firms	britain
ended	increasing	benefited	oil	unit	group	value	spokesman	suez-med
stores	technology	increasing	carbon	subsidiary	bringing	currencies	japan's	deadlines
store	defense	mirrored	cubic	completed	receipts	currency	officials	geonex
retail	process	umar.1	energy	work	zayre	rates	quoted	occurs
weeks	licensing	voluntary	fuel	dlr	add	world	saying	explains
customers	better	chronar	million	received	semi-official	coins	products	le

TABLE I: Topics and their most probable word lists for Source-LDA, IR-LDA, and CTM.

crawled and the words in the topic were counted, forming their respective distributions. Querying Wikipedia resulted in 80 distinct topics as our superset for the knowledge source. Out of the 80 crawled available topics, only 49 topics appear in the 2,000 document corpus. This represents the ideal conditions in which Source-LDA is to be applied; that of a corpus which a significant portion of tokens are generated from a subset of a larger and relatively easy to obtain topic set. For all models, a symmetric Dirichlet parameter of $50/T$ (where T is the number of topics) and $200/V$ (where V is the size of the vocabulary) was used for α and β respectively. For Source-LDA, μ and σ were determined by experimentally finding a local minimum value of perplexity which resulted from the parameter values of 0.7 for μ and 0.3 for σ. The bag of words used in the CTM were taken from the top 10,000 words by frequency for each topic. The models showed good convergence after 1,000 iterations. After sampling was complete for LDA, the resulting topic-to-word distribution was mapped using an information retrieval (IR) approach. The IR approach was to use cosine similarity of documents mapped to term frequency-inverse document frequency (TF-IDF) vectors with TF-IDF weighted query vectors formed from the top 10 words per topic.

2) Experimental Results: After the LDA model converged, we label the topics using the IR approach described above (we referred to this topic labeling method as IR-LDA). Given similar labels from the models it is an intuitive approach to compare the word assignments to each topic model. Example comparisons are shown in Table I. The label assignments generated from Source-LDA show a more accurate assignment of labels to topics than both IR-LDA and CTM. IR-LDA appears to suffer from mixing of different concepts into a single topic, for example with the topic "Japanese Yen," the topic assignments could possibly be the combination of "Japanese Yen" and "Japanese Exports". The CTM seems to assign more weight to less important words. One approach to rectify this problem for CTM is to use a smaller number of words for the bag of words, but this leads to significant dropout and no labeled topics are passed through. Source-LDA is more consistent with the meaning of the topic as opposed to what words you may find when talking about this topic, which can be generally applied to many concepts.
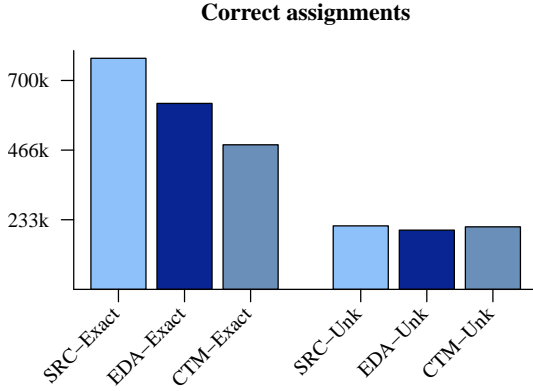
D. Wikipedia Corpus

A comparison of Source-LDA against EDA, and CTM is made using a corpus generated using a known knowledge source corresponding to medical topics extracted from Med-linePlus (a consumer-friendly medical dictionary) [27]. We evaluate the strength of Source-LDA under different models proposed in Section 3 using the metrics of classification accuracy, JS divergence and Pointwise mutual information (PMI).

PMI is an established evaluation of learned topics which takes as input a subset of the most popular tokens comprising a topic and determines the frequency of all pairs in the subset occurring at a given input distance from each other in the corpus. The more that these pairs occur close to each other then the better the learned topics. PMI differs from the JS divergence evaluation for this experiment in that PMI will tell us how good our topics are where as the JS divergence will tell us how good our distribution over topics for each document is.

1) Experimental Setup: A corpus of Wikipedia vocabulary articles was generated by following the steps of the generative model for Source-LDA, where the chosen K topics are a subset of a larger collection of Wikipedia topics. The topics consisted of 578 Wikipedia articles representing the collection of topic labels from MedlinePlus. The number of topics (K) was given as 100, chosen from an entire collection of 578 topics (B), the number of documents (D) was given as 2000 and the average document word count (D_{avg}) as 500, μ and σ were set to 5.0 and 2.0 for the bijective evaluation 0.7 and 0.3 for the Source-LDA model respectively. After these 2000 documents were generated the topic assignments were recorded and used as the ground truth measurement. The word assignments were used as the corpus and the different topic models were applied to these documents. The first round of topic models consisted of comparing Source-LDA, EDA, and CTM. For Source-LDA μ and σ were set to match that of the generative model. For all models, a symmetric Dirichlet parameter of $50/T$ and $200/V$ was used for α and β respectively. After convergence of the models they were evaluated against the ground truth measurement. In the second round of experiments each topic model was run under the bijective model, that is they only considered topics which were used in the ground truth assignments.

To compare Source-LDA against LDA using PMI, a separate corpus was generated under the bijective model with K, B, D, D_{avg}, μ, and σ set to 100, 578, 200, 300, 1.0 and 0.0 respectively. The parameters for Source-LDA followed the generative model and all other parameters are the same as the previous experiments. After 1000 iterations the top 10 words
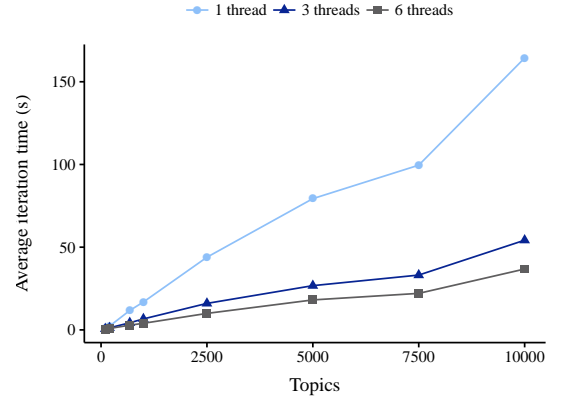
Correct assignments

(a)

Jensen−Shannon divergence
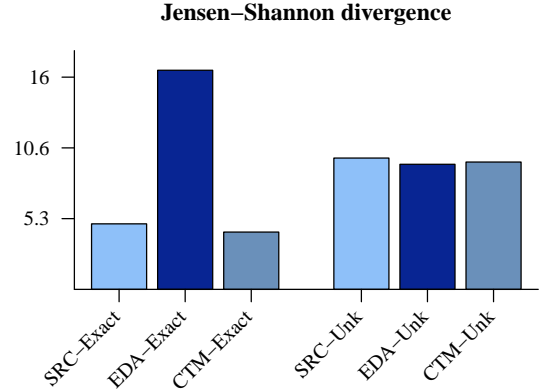
(b)
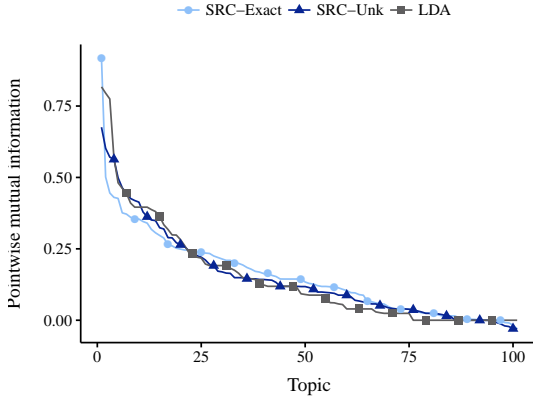
(c)

(d)

Fig. 8: Results showing the number of correct topic assignments in the mixed model and bijective model (a) and sum total of the JS divergences of θ in the mixed and bijective models (b). Sorted PMI analysis for a Wikipedia generated corpus inferred by the exact bijective model and mixed model is shown by (c). Performance benchmarking is given in (d).

given for each topic were used in the PMI assessment.

2) Experimental Results: The topic assignments for each token in the corpus were recorded for all models and the results compared against each other. Since we know a priori the correct topic assignment for each token we use the number of correct topic assignments to be an appropriate measure of classification accuracy. In Figure 8, all topic models run under the full Source-LDA model are tagged with an "Unk" label, and likewise topic models run under the bijective model are tagged with "Exact". The overall number of correct topic assignments for each model are shown in Figure 8(a) for the mixed model the bijective model. Note that in evaluations where the ground truth is known, classification accuracy is a much better determination of the goodness of a model than log likelihood maximizations such as perplexity and therefore we do not evaluate the model using perplexity. As expected the Source-LDA model (SRC-Unk and SRC-Exact) had the best results amongst all other topic models for classification accuracy.

In the second analysis the topic to document distributions were analyzed using JS Divergence. The results again show the Source-LDA model to be effective in accurately mapping topics to documents whether or not the topics used in the generative model are unknown or a known set of topics as

shown in (Figure 8(b)).

The PMI analysis detailed by Figure 8(c) show that by PMI, Source-LDA provides a better mapping of labels to topics on average as the area under the curve for Source-LDA is larger than that of LDA in the Wikipedia generated corpus. This is an encouraging result, even though the differences are not large, since LDA is a function of topic proximity in a document and word frequency in a topic, whereas Source-LDA is a function of the same plus the likelihood of a word being in an augmented source distribution.

E. Performance Benchmarking

To show the performance gains used by the parallel sampling algorithm and experiment was set up to generate topics randomly from a given vocabulary. The corpus was generated using the same parameters as in Section 4(B) but with B ranging from 100 to 10000. The benchmarking is visualized by Figure 8(d). It clearly demonstrates that Source-LDA is linearly scalable and easily parallelized.

V. RELATED WORK

Much existing literature exists related to the proposed approach in this paper. These methods are mainly extensions

of LDA, and add to the original model by introducing enhancements such as topic labeling, integration with contextual information and hierarchical modeling.

A. Topic Labeling

In the early research stage, labels were often generated by hand [28]–[31]. Though manual labeling may generate more understandable and accurate semantics of a topic, it costs a lot of human effort and it is prone to subjectivity [32]. For example, in the most conventional LDA model, topics are interpreted by selecting the top words in the distribution [1], [28], [32], [33]. The Topics over Time (TOT) model implements continuous time stamps with each topic [32]. The model has been applied in three kinds of datasets, and results show more accurate topics and better timestamp predictions. However, the interpretation of topics is manual and post-hoc labeling can be time-consuming and subjective.

Mei et al. proposed probabilistic approaches to automatically interpreting multinomial topic models objectively. The intuition of this algorithm was to minimize the semantic distance between the topic model and the label. To this end, they extracted candidate labels from noun phrases chunked by an NLP Chunker and most significant 2-grams. Then they ranked labels to minimize Kullback-Leibler divergence and maximize mutual information between a topic model and a label. The approach achieved the automatic interpretation of topics, but available candidate labels were limited to phrases inside documents.

Lau et al came up with an automatic topic label generation method which obtains candidate labels from Wikipedia articles containing the top-ranking topic terms, top-ranked document titles, and sub-phrases. To rank those candidates topic labels, they used different lexical measurements, such as point-wise mutual information, Student's t-test, Dice's coefficient and the log likelihood ratio [34]. Supervised methods like support vector regression were also applied in the ranking process. Results showed that supervised algorithm outperforms unsupervised baseline in all four corpora.

In previous approaches, topics were treated individually and relation among topics was not considered. Mao et al created hierarchical descriptor for topics, and results proved that inner-topic relation could increase the accuracy of topic labels [35]. Hulpus et al proposed a graph-based approach for topic labeling [36]. In Yashar Mehdad's work, they built an entailment graph over phrases. Based on that, they then aggregated relevant phrases by generalization and merging [37].

Conceptual labeling is an approach to generate a minimum sized set of labels that best describe a bag of words which includes topics generated from topic modeling [38]. Concepts used in the topic labeling are taken from a semantic network and deemed appropriate using the metric Minimum Description Length. This approach is applied after topic modeling and represents an effective way of labeling topics over existing approaches.

B. Supervised Labeling

Supervised Latent Dirichlet Allocation (sLDA) is a supervised approach to labeling topics [14]. The approach includes a response variable into the LDA model to obtain latent topics that potentially provide an optimal prediction for the response variable of a new unlabeled document. This approach requires,

during training, the manual input of individual topic labels and is constrained to permitting one label per topic.

Similar to sLDA is Discriminative LDA (DiscLDA) which attempts to solve the same problem as sLDA, but differs in the approach [15]. The differing approach was centered around introducing a class-dependent linear transformation on the topic mixture proportions. This transformation matrix was learned through a conditional likelihood criterion. This method has the benefit of both reducing the dimension of documents in the corpus and labeling the lower dimension documents.

Both sLDA and DiscLDA only allow for a supervised input set that label a single topic. An approach that allows for multiple labels in a topic is given by Labeled LDA (L-LDA) [16]. This model differs in the generation of multinomial distribution theta over the topics in the model. The scaling parameter is then modified by a label projection matrix to restrict the distribution to those topics considered most relevant to the document.

C. Contextual Integration

An existing approach that takes into account concepts supplied by prior sources requires a manual input set of relevant terms [39]. In the topic model then these concepts are applied to the assignment of topics to a token in a document. Alongside this concept topic modeling a hierarchical method can also be used to incorporate concepts into a hierarchical structure. This work shows the utility of bringing in prior knowledge into topic modeling.

An approach that integrates Wikipedia information into the topic modeling differs than the supervised approach by only requiring an existing Wikipedia article [7]. The assumption in this work is that in the generative process the topics are selected from the Wikipedia word distributions. The results show that Wikipedia articles can be used as effective topics in topic modeling.

Wikipedia again was shown as a basis for topic modeling, albeit for a tangential approach, entity disambiguation [7]. The approach involved topic modeling as a way of annotating entities in text. This involved the use of a large dataset of topics so efficient methods were introduced. Experiments against a public dataset resulted in a state of the art performance.

VI. Conclusion

We have described in this paper a novel methodology for semi-supervised topic modeling with meaningful labels, as well as provided parallel algorithms to speed up the inference process. This methodology uses prior knowledge sources to influence a topic model in order to allow the labels from these external sources to be used for topics generated over a corpus of interest. In addition, this approach results in more meaningful topics generated based on the quality of the external knowledge source. We have tested our methodology against the Reuters-21578 newswire collection corpus for labeling and Wikipedia as external knowledge sources. The analysis of the quality of topic models using PMI show the ability of Source-LDA to enhance existing topic models.

References

[1] D. M. Blei *et al.*, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] R. S. Margalit *et al.*, "Electronic medical record use and physician-patient communication: an observational study of Israeli primary care encounters," *Patient Education and Counseling*, vol. 1, pp. 131–141, 2006.

[3] C. W. Arnold *et al.*, "Clinical case-based retrieval using latent topic analysis," *AMIA Annual Symposium Proceedings*, vol. 2010, p. 26, 2010.

[4] H. Bisgin *et al.*, "Mining FDA drug labels using an unsupervised learning technique - topic modeling," *BMC Bioinformatics*, vol. 12, no. S-10, p. S11, 2011.

[5] W. Speier, M. K. Ong, and C. W. Arnold, "Using phrases and document metadata to improve topic modeling of clinical reports," *Journal of Biomedical Informatics*, vol. 61, pp. 260–266, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2016.04.005

[6] C. C. others, "Text modeling using unsupervised topic models and concept hierarchies," *CoRR*, vol. abs/0808.0973, 2008.

[7] J. A. Hansen *et al.*, "Probabilistic explicit topic modeling using wikipedia," in *Language Processing and Knowledge in the Web - 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25-27, 2013. Proceedings*, 2013, pp. 69–82.

[8] J. Jagarlamudi *et al.*, "Incorporating lexical priors into topic models," in *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, 2012, pp. 204–213.

[9] T. P. Minka, "Bayesian inference, entropy, and the multinomial distribution," 2000.

[10] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl. 1, pp. 5228–5235, Apr. 2004.

[11] W. M. Darling, "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 642–647.

[12] T. Griffiths, "Gibbs sampling in the generative model of latent dirichlet allocation," 2002.

[13] M. W. Beck, "Average dissertation and thesis length," https://github.com/fawda123/diss_proc, 2014.

[14] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, 2007, pp. 121–128.

[15] S. Lacoste-Julien *et al.*, "Disclda: Discriminative learning for dimensionality reduction and classification," in *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, 2008, pp. 897–904.

[16] D. Ramage *et al.*, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2009, pp. 248–256.

[17] Y. Wang *et al.*, "PLDA: parallel latent dirichlet allocation for large-scale applications," in *Algorithmic Aspects in Information and Management, 5th International Conference, AAIM 2009, San Francisco, CA, USA, June 15-17, 2009. Proceedings*, 2009, pp. 301–314.

[18] D. Newman *et al.*, "Distributed inference for latent dirichlet allocation," in *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, 2007, pp. 1081–1088.

[19] I. Porteous *et al.*, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, 2008, pp. 569–577.

[20] G. E. Blelloch, "Prefix sums and their applications," Synthesis of Parallel Algorithms, Tech. Rep., 1990.

[21] H. M. Wallach *et al.*, "Rethinking LDA: why priors matter," in *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, 2009, pp. 1973–1981. [Online]. Available: http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter

[22] H. M. Wallach *et al.*, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, 2009, pp. 1105–1112. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553515

[23] G. Heinrich, "Parameter estimation for text analysis," *University of Leipzig, Tech. Rep.*, 2008.

[24] J. Kang *et al.*, "Transfer topic modeling with ease and scalability," in *Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012.*, 2012, pp. 564–575. [Online]. Available: http://dx.doi.org/10.1137/1.9781611972825.49

[25] J. Chang *et al.*, "Reading tea leaves: How humans interpret topic models," in *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, 2009, pp. 288–296. [Online]. Available: http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models

[26] C. W. Arnold, A. Oh, S. Chen, and W. Speier, "Evaluating topic model interpretability from a primary care physician perspective," *Computer Methods and Programs in Biomedicine*, vol. 124, pp. 67–75, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2015.10.014

[27] "Medlineplus [internet]," https://www.nlm.nih.gov/medlineplus/.

[28] Q. Mei *et al.*, "Automatic labeling of multinomial topic models," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, 2007, pp. 490–499.

[29] Q. Mei *et al.*, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, 2006, pp. 533–542.

[30] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005*, 2005, pp. 198–207.

[31] Q. Mei and C. Zhai, "A mixture model for contextual text mining," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, 2006, pp. 649–655.

[32] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, 2006, pp. 424–433.

[33] J. H. Lau *et al.*, "Automatic labelling of topic models," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, 2011, pp. 1536–1545.

[34] P. Pecina, "Lexical association measures and collocation extraction," *Language Resources and Evaluation*, vol. 44, no. 1-2, pp. 137–158, 2010.

[35] X. Mao *et al.*, "Automatic labeling hierarchical topics," in *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, 2012, pp. 2383–2386.

[36] I. Hulpus *et al.*, "Unsupervised graph-based topic labelling using dbpedia," in *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, 2013, pp. 465–474.

[37] Y. Mehdad *et al.*, "Towards topic labeling with phrase entailment and aggregation," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, 2013, pp. 179–189.

[38] X. Sun *et al.*, "On conceptual labeling of a bag of words," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2015, pp. 1326–1332. [Online]. Available: http://ijcai.org/Abstract/15/191

[39] M. Steyvers *et al.*, "Combining background knowledge and learned topics," *topiCS*, vol. 3, no. 1, pp. 18–47, 2011.