Nested Hierarchical Dirichlet Process for Nonparametric Entity-Topic Analysis

Priyanka Agrawal[1]*, Lavanya Sita Tekumalla[2]*, and Indrajit Bhattacharya[1]

[1] IBM India Research Lab,
priyanka.svnit@gmail.com, indrajitb@gmail.com,
[2] Indian Institute of Science,
lavanya@csa.iisc.ernet.in

Abstract. The Hierarchical Dirichlet Process (HDP) is a Bayesian nonparametric prior for grouped data, such as collections of documents, where each group is a mixture of a set of shared mixture densities, or topics, where the number of topics is not fixed, but grows with data size. The Nested Dirichlet Process (NDP) builds on the HDP to cluster the documents, but allowing them to choose only from a set of specific topic mixtures. In many applications, such a set of topic mixtures may be identified with the set of entities for the collection. However, in many applications, multiple entities are associated with documents, and often the set of entities may also not be known completely in advance. In this paper, we address this problem using a nested HDP (nHDP), where the base distribution of an outer HDP is itself an HDP. The inner HDP creates a countably infinite set of topic mixtures and associates them with entities, while the outer HDP associates documents with these entities or topic mixtures. Making use of a nested Chinese Restaurant Franchise (nCRF) representation for the nested HDP, we propose a collapsed Gibbs sampling based inference algorithm for the model. Because of couplings between two HDP levels, scaling up is naturally a challenge for the inference algorithm. We propose an inference algorithm by extending the direct sampling scheme of the HDP to two levels. In our experiments on two real world research corpora, we show that, even when large fractions of author entities are hidden, the nHDP is able to generalize significantly better than existing models. More importantly, we are able to detect missing authors at a reasonable level of accuracy.

1 Introduction

Dirichlet Process mixture models [1] allow for nonparametric or infinite mixture modeling, where the number of densities or mixture components is not fixed ahead of time, but grows (slowly) with the number of data items. They do so by using as a prior the Dirichlet Process (DP), which is a distribution over distributions, and has the additional property that draws from it are discrete (w.p. 1) with infinite support [1, 6]. However, many applications require joint

* The first two authors have contributed equally to the paper

analysis of groups of data, such as a collection of text documents, where the mixture components, or topics (as they are called for text data), are shared across the documents. This calls for a coupling of multiple DPs, one for each document, where the base distribution is discrete, and shared. The hierarchical Dirichlet Process (HDP) [16] does so by placing a DP prior on a shared base distribution, so that the model now has two levels of DPs. The HDP has since been used extensively as a prior for non-parametric modeling of text collections. The popular LDA model [3] may be considered as a parametric restriction of the HDP mixture model.

The HDP mixture model (and LDA) belongs to the family of admixture models [5], where each composite data item or group gets assigned to a mixture over the mixture components or topics. While this adds more flexibility to the groups of data items, the ability to cluster groups is lost, since each group now has a distinct mixture of topics associated with it. This additional capability is desired in many applications, such as analysis of patient profiles in hospitals [13], where the hospitals need to be clustered in addition to shared grouping of patients in individual hospitals. Alternatively, imagine a corpus containing descriptions related to entities, such as a shared set of researchers who have authored a large body of scientific literature, or a shared set of personalities discussed across news articles, such that each entity can be represented as a mixture of topics. Here, topic mixtures, corresponding to entities, are required to be shared across data groups or documents, in addition to the topics themselves. This can be captured using the nested DP (nDP) [13], which has a DP corresponding to each group, which are coupled through the same base distribution, which is a DP itself, unlike being DP distributed, as in the HDP. This results in a distribution over distributions over distributions, unlike the HDP and the DP, which are distributions over distributions. The nDP can be imagined as first creating a discrete set of mixtures over topics, each mixture representing an entity, and then choosing exactly one of these entities for each document. In this sense, the nDP is a mixture of admixtures.

One major shortcoming of the nDP for entity analysis is the restrictive assumption of a single entity being associated with a document. In research papers, multiple authors are associated with any document, and any news article typically discusses multiple people, organizations etc. This requires each document to have a distribution over entities. In other words, we need a model that is an admixture of admixtures. The Author-Topic Model (ATM) [14], which models authors associated with documents, belongs to this class, but is restrictive in that it requires the authors to be observed for documents, and also assumes the number of topics to be known.

In this paper, we address the problem of nonparametric modeling of entities and topics, where the number of topics is not known in advance, and additionally the set of entities for each document is either partly or completely unknown. For this, we propose the nested HDP (nHDP), where the base distribution of an HDP is itself an HDP. This belongs to the same class as the nDP, since it specifies a distribution over distributions (entities) over distributions (topics). However, unlike the nDP, it first creates a discrete set of entities, and models each group as a mixture over these entities. To the best of our knowledge, ours is the first entity-topic model that is nonparametric in both entities and topics. The Author Topic Model falls out as a parametric version of this model, when the entity set is observed for each document, and the number of topics is fixed.

For inference using the nHDP, we propose the nested CRF (nCRF), which extends the Chinese Restaurant Franchise (CRF) analogy of the HDP to two levels by integrating out the two levels of HDPs. However, due to strong coupling between the CRF layers, inference using the nCRF poses computational challenges. We use a direct sampling scheme, based on that for the HDP, where the entity and topic indexes are directly sampled, based on the counts of table assignments and stick-breaking weights at the two levels. Using experiments over publication datasets involving author entities from NIPS and DBLP, we show that the nHDP generalizes better under different levels of available author information. More interestingly, the model is able to detect authors completely hidden in the entire corpus with reasonable accuracy.

2 Related Work

In this section, we review existing literature on Bayesian nonparametric modeling and entity-topic analysis.

Bayesian Nonparametric Models: We review the Dirichlet Process (DP) [6], the Hierarchical Dirichlet Process (HDP) [16] and the nested Dirichlet Process (nDP) [13] in detail in the Sec. 3.

The MLC-HDP[17] is a 3-layer model proposed for human brain seizures data. The 2-level truncation of the model is closely related to the HDP and the nDP. Like the HDP, it shares mixture components across groups (documents) and assigns individual data points to the same set of mixtures, and like the nDP it clusters each of the groups or documents using a higher level mixture. In other words, this is a nonparametric mixture of admixtures, while our proposed nested HDP is a nonparametric admixture of admixtures.

The nested Chinese Restaurant Process (nCRP) [2] extends the Chinese Restaurant Process analogy of the Dirichlet Process to an infinitely-branched tree structure over restaurants to define a distribution over finite length paths of trees. This can be used as a prior to learn hierarchical topics from documents, where each topic corresponds to a node of this tree, and each document is generated by a random path over these topics. An extension to this model, also called the nested HDP, has recently been proposed on Arvix [11]. In the spirit of the HDP, which has a top level DP and providing base distributions for document specific DPs, this model has a top level nCRP, which becomes the base distribution for document specific nCRPs. In contrast, our model has nested HDPs, in the sense that one HDP directly serves as the base distribution for another HDP, like in the nested DP [13], where one DP serves as the base distribution for another DP. This parallel with the nested DP motivates the nomenclature of our model as the nested HDP.

Entity-Topic Models: Next, we briefly review prior work on simultaneously modeling of entities and topics in documents. The literature mostly contains parametric models, where the number of topics and entities are known ahead of time. The LDA model [3] is the most popular parametric topic model, and has a distribution θ_d over T topics for each document, and the topic label z_{di} for each word in the document is sampled *iid* from θ_d. The author-topic model (ATM)[14] extends the LDA to capture *known* authors A_d of each document. Each author now has his own distribution π_a over topics K, and the words are generated by first sampling one of the known authors uniformly, followed by sampling a topic from the topic distribution of that author:

$$\begin{aligned} \phi_k \sim Dir(\beta),\ k = 1 \dots T;\ \ \pi_a \sim Dir(\alpha),\ a = 1 \dots A \\ a_{di} \sim \pi_d \equiv U(A_d);\ z_{di} \sim \theta_{a_{di}};\ w_{di} \sim Mult(\phi_{z_{di}}) \end{aligned} \tag{1}$$

The Author Recipient Topic model[9] distinguishes between sender and recipient entities and learns the topics and topic distributions of sender-recipient pairs. Newman et. al[10] analyze entity-topic relationships from textual data containing entity words and topic words, which are pre-annotated. The Entity Topic Model[8] proposes a generative model which is parametric in both entities and topics and assumes observed entities for each document.

There has been very little work on nonparametric entity-topic modeling, which would enable discovery of entities in settings where entities are partially or completely unobserved in documents. The Author Disambiguation Model[4] is a nonparametric model for the author entities along with topics. Primarily focusing on author disambiguation from noisy mentions of author names in documents, this model treats entities and topics symmetrically, generating entity-topic pairs from a DP prior. Contrary to this approach, our model treats the entity as a distribution over topics, thus explicitly modeling the fact that authors of documents have preferences over specific topics.

3 Background

Consider a setting where observations are organized in groups. Let x_{ji} denote the i^{th} observation in j^{th} group. For a corpus of documents, x_{ji} is the i^{th} word occurrence in the j^{th} document. In the context of this paper, we will use group synonymously with document, data item with word in a document. We assume that each x_{ji} is independently drawn from a mixture model and has a mixture component parameterized by a *factor*, say θ_{ji}, representing a topic, associated with it. For each group j, let the associated factors $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \dots)$ have a prior distribution G_j. Finally, let $F(\theta_{ji})$ denote the distribution of x_{ji} given factor θ_{ji}. Therefore, the generative model is given by

$$\theta_{ji}|G_j \sim G_j;\ x_{ji}|\theta_{ji} \sim F(\theta_{ji}),\ \forall j, i \tag{2}$$

The central question in analyzing a corpus of documents is the parametrization of the G_j distributions — what parameters to share and what priors to place on them. We start with the Dirichlet Process that considers each of the G_j distributions in isolation, then the Hierarchical Dirichlet Process that ensures sharing of atoms among the different G_js, and finally the nested Dirichlet Process that additionally clusters the groups by ensuring that all the G_js are not distinct.

Dirichlet Process: Let $(\Theta, \mathcal{B})$ be a measurable space. A Dirichlet Process (DP) [6, 1] is a measure over measures G on that space. Let G_0 be a finite measure on the space. Let α_0 be a positive real number. We say that G is DP distributed with concentration parameter α and base distribution G_0, written $G \sim$DP(α_0, G_0), if for any finite measurable partition $(A_1, \ldots, A_r)$ of Θ, we have

$$(G(A_1), \ldots G(A_r)) \sim Dir(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)).$$

The *stick-breaking representation* provides a constructive definition for samples drawn from a DP. It can be shown [15] that draw G from $DP(\alpha_0, G_0)$ can be written as

$$\phi_k \overset{iid}{\sim} G_0,\ k = 1 \ldots \infty; \quad w_i \sim Beta(1, \alpha_0);\ \beta_i = w_i \prod_{j=1}^{i-1}(1 - w_j)$$
$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \tag{3}$$

where the atoms ϕ_k are drawn independently from G_0 and the corresponding weights $\{\beta_k\}$ follow a stick breaking construction. This is also called the GEM distribution: $(\beta_k)_{k=1}^{\infty} \sim$ GEM(α_0). The stick breaking construction shows that draws from the DP are necessarily discrete, with infinite support, and the DP therefore is suitable as a prior distribution on mixture components for 'infinite' mixture models. Subsequently, $\{\theta_{ji}\}$ are drawn from each G_j. When drawing of $\{\theta_{ji}\}$ is followed by draws $\{x_{ji}\}$ according to Eqn. 2, the model is known as the Dirichlet Process mixture model [6].

Another commonly used perspective of the DP is the *Chinese Restaurant Process* (CRP) [12], which shows that DP tends to clusters the observations. Let $\{\theta_i\}$ denote the sequence of draws from G, and let $\{\phi_k\}$ be the atoms of G. The CRP considers the predictive distribution of the i^{th} draw θ_i given the first $i-1$ draws $\theta_1 \ldots \theta_{i-1}$ after integrating out G:

$$\theta_i | \theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i - 1 + \alpha_0} \delta_{\phi_k} + \frac{\alpha_0}{i - 1 + \alpha_0} G_0$$

where $m_k = \sum_{i'=1}^{i-1} \delta(\theta_{i'}, \phi_k)$. The above conditional may be understood in terms of the following restaurant analogy. Consider an initially empty 'restaurant' that can accommodate an infinite number of 'tables'. The i^{th} 'customer' entering the restaurant chooses a table θ_i for himself, conditioned on the seating arrangement of all previous customers. He chooses the k-th table with probability proportional to m_k, the number of people already seated at the table, and with probability proportional to α_0, he chooses a new (currently unoccupied) table. Whenever a new table is chosen, a new 'dish' ϕ_k is drawn ($\phi_k \sim G_0$) and associated with the table. The CRP readily lends itself to sampling-based inference strategies for the DP.

Hierarchical Dirichlet Process: Now reconsider our grouped data setting. If each G_j is drawn independently from a DP, then w.p. 1 the atoms $\{\phi_{jk}\}_{k=1}^{\infty}$ for each G_j are distinct. This would mean that there is no shared topic across documents, which is undesirable. The Hierarchical Dirichlet Process (HDP) [16] addresses this problem by modeling the base distribution G_0 of the DP prior in turn as DP distributed. Since draws from a DP are discrete, this ensures that the same atoms $\{\phi_k\}$ are shared across all the G_js. Specifically, given a distribution H on the space $(\Theta, \mathcal{B})$ and positive real numbers $(\alpha_j)_{j=1}^{M}$ and γ, we denote as HDP$(\boldsymbol{\alpha}, \gamma, H)$ the following generative process:

$$\begin{aligned} &G_0|\gamma, H \sim DP(\gamma, H) \\ &G_j|\alpha_j, G_0 \sim DP(\alpha_j, G_0) \qquad \forall j. \end{aligned} \tag{4}$$

When this is followed by generation of $\{\theta_{ji}\}$ and $\{x_{ji}\}$ as in Eqn. 2, we get the *HDP mixture model.*

Using the stick-breaking construction, the global measure G_0 distributed as Dirichlet process can be expressed as $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$, where the topics ϕ_k are drawn from H independently ($\phi_k \sim H$) and $\{\beta_k\} \sim$ GEM(γ) represent 'global' popularities of these topics. Since G_0 has as its support the topics $\{\phi_k\}$, each group-specific distribution G_j necessarily has support at these topics, and can be written as follows:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}; \quad (\pi_{jk})_{k=1}^{\infty} \sim \mathrm{DP}(\alpha_j, \boldsymbol{\beta}) \tag{5}$$

where $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty}$ denotes the topic popularities for the jth group.

Analogous to the CRP for the DP, the Chinese Restaurant Franchise provides an interpretation of predictive distribution for the next draw from an HDP after integrating out the G_js and G_0. Let $\{\theta_{ji}\}$ denote the sequence of draws from each G_j, $\{\psi_{jt}\}$ the sequence of draws from G_0, and $\{\phi_k\}$ the sequence of K draws from H. Then the conditional distribution of θ_{ji} given $\theta_{j1}, \ldots, \theta_{j,i-1}$ and G_0, after integrating out G_j is as follows:

$$\theta_{ji}|\theta_{j1}, \ldots, \theta_{j,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i-1+\alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \tag{6}$$

where $n_{jtk} = \sum_{i'=1}^{i-1} \delta(\theta_{ji'}, \psi_{jt}) \delta(\psi_{jt}, \phi_k)$, $m_{jk} = \sum_t \delta(\psi_{jt}, \phi_k)$ and dots indicate marginal counts. As G_0 is also distributed according to a Dirichlet Process, we can integrate it out similarly to get the conditional distribution of ψ_{jt}:

$$\psi_{jt}|\psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{jt-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma} H \tag{7}$$

These equations may be interpreted using a two-level restaurant analogy. Consider a set of restaurants, one corresponding to each group. Customers entering each of the restaurants select a table θ_{ji} according a group specific CRP (Eqn

6). The restaurants share a common menu of dishes $\{\phi_k\}$. Dishes are assigned to the tables of each restaurant according to another CRP (Eqn 7). Let t_{ji} be the (table) index of the element of $\{\psi_{jt}\}$ associated with θ_{ji}, and let k_{jt} be the (dish) index of the element of $\{\phi_k\}$ associated with ψ_{jt}. Then the two conditional distributions above can also be written in terms of the indexes $\{t_{ji}\}$ and $\{k_{jt}\}$ instead of referring to the distributions directly. If we draw ψ_{jt} by choosing a summation term, we set $\psi_{jt} = \phi_k$ and let $k_{jt} = k$ for the chosen k. If the second term is chosen, we increment K by 1 and draw $\phi_K \sim H$ and set $\psi_{jt} = \phi_K$ and $k_{jt} = K$. This CRF analogy leads to efficient Gibbs sampling-based inference strategies for the HDP mixture model [16].

Nested Dirichlet Process: In other applications of grouped data, we may additionally be interested in clustering the data groups themselves. For example, when analyzing patient records in hospitals, we may want to cluster the hospitals as well in terms of the profiles of patients coming there. The HDP cannot do this, since each group specific mixture G_j is distinct. This problem is addressed by the nested Dirichlet Process [13], which first defines a set $\{G'_k\}_{k=1}^{\infty}$ of distributions over an infinite support:

$$G'_k = \sum_{l=1}^{\infty} w_{lk}\delta_{\theta'_{lk}},\ \theta'_{lk} \sim H,\ \{w_{lk}\}_{l=1}^{\infty} \sim GEM(\gamma) \tag{8}$$

and then draws the group specific distributions G_j from a mixture over these:

$$G_j \sim G_0 \equiv \sum_{k=1}^{\infty} \pi_k \delta_{G'_k},\ \{\pi_k\} \sim GEM(\alpha)$$

We denote the generation process as $\{G_j\} \sim nDP(\alpha, \beta, H)$. The process ensures non-zero probability of different groups selecting the same G'_k, leading to clustering of the groups. Using Eqn. 3, the draws $\{G_j\}$ can be characterized as:

$$G_j \sim G_0,\ G_0 \sim DP(\alpha, DP(\gamma, H)) \tag{9}$$

where the base distribution of the outer DP is in turn another DP, unlike the HDP where it is DP distributed. Thus the nDP can be viewed as a distribution on the space of distributions on distributions.

Given this characterization of the nDP, it appears to be useful for restricted entity-topic analysis, where we additionally want to label each document with a single entity from a countable set, with each entity associated with a distribution over topics. However, note that the support $\{\theta'_{lk}\}$ of each G'_k in Eqn 8 is distinct, which implies that different entities do not share any topics. Further, we would like to associate a distribution over entities for each document. This makes the nDP unsuitable even for such restricted entity-topic analysis.

4 Nonparametric Entity-Topic Analysis

We now present our nested Hierarchical Dirichlet Process (nHDP) model for nonparametric entity-topic analysis. We first present the model where each group or document is associated with a single entity, then extend it for multiple entities.

Single-Entity Documents: Recall that the nDP is unsuitable for entity-topic analysis, since the entity distributions do not share topic atoms. This can be modified by first creating a set of entity distributions $\{G_{k'}\}_{k'=1}^{\infty}$ such that they share atoms. One way to do this is to follow the HDP construction:

$$G_{k'} \sim HDP(\{\alpha_{k'}\}, \gamma, H) \tag{10}$$

This can be followed by drawing each group specific distribution from a mixture over the $G_{k'}$s:

$$G'_j \sim G'_0 \equiv \sum_{k'=1}^{\infty} \beta'_{k'} \delta_{G_{k'}},\ \beta' \sim GEM(\gamma') \tag{11}$$

Using Eqn. 3, we observe that $G'_0 \sim DP(\gamma', HDP(\ \{\alpha_{k'}\}, \gamma, H))$. Observe the relationship with the nDP (Eqn. 9). Like the nDP, this also defines a distribution over the space of distributions on distributions. But, instead of a DP base distribution for the outer DP, we have achieved sharing of atoms using a HDP base distribution. We will write $G'_j \sim$ DP-HDP$(\gamma', \{\alpha_{k'}\}, \gamma, H)$.

Sampling G'_j may be imagined as choosing the entity for the j^{th} document. As before, G'_j can now be used as prior for sampling topics $\{\theta_{ji}\}$ for individual words in document j, followed by sampling of the words themselves, using Eqn 2. We will call this the DP-HDP mixture model.

Nested HDP for Multi-Entity Documents: The DP-HDP model above associates a single distribution over topics G'_j with the j^{th} document, and the topic θ_{ji} for each word x_{ji} in the document is drawn from G'_j. In the context of entity-topic analysis, this means that a single entity is associated with a document, and words are drawn from the preferred topics of this entity. However, many applications, such as analyzing entities in news articles and authors from scientific literature, require associating multiple entities with each document, and each word in the document is drawn from a preferred topic of one of these entities. In this section, we extend the earlier model for multi-entity documents.

As before, we first create a set of distributions $\{G_{k'}\}_{k'=1}^{\infty}$ over the same set of (topic) distributions $\{\phi_k\}_{k=1}^{\infty}$ ($\phi_k \sim H$) by drawing independently from an HDP, and creating a global mixture over them:

$$G_{k'} \sim HDP(\{\alpha_{k'}\}, \gamma, H); \quad \beta' \sim GEM(\gamma'); G'_0 \equiv \sum_{k'=1}^{\infty} \beta'_{k'} \delta_{G_{k'}}$$

This may be interpreted as creating a countable set of entities by defining topic preferences (distributions over topics) for each of them, and then defining a 'global popularity' of the entities. Earlier, for single entity documents, the only entity was sampled from this global popularity. Now, we define a document-specific local popularity for entities, derived from this global popularity:

$$G'_j \equiv \sum_{k'=1}^{\infty} \pi'_{jk'} \delta_{G_{k'}},\ \{\pi'_{jk'}\} \sim DP(\alpha'_j, \beta') \tag{12}$$

Now, sampling each factor θ_{ji} in document j is preceded by choosing an entity $\theta'_{ji} \sim G'_j$ by sampling according to local entity popularity. Note that $P(\theta'_{ji} = G_{k'}) = \pi'_{jk'}$.

Using the stick breaking definition of the HDP in Eqn. 5, we can see that G'_j is drawn from a HDP. The base distribution of that HDP has to be the distribution from which the atoms $\{G_{k'}\}$ are drawn, which is again an HDP. Therefore, we can write:

$$\theta'_{ji} \sim G'_j \sim \mathrm{HDP}(\{\alpha'_j\}, \gamma', \mathrm{HDP}(\{\alpha_{k'}\}, \gamma, H)) \tag{13}$$

We refer to the two relevant HDPs as the inner and outer HDPs. Observing the parallel with the nDP definition in Eqn. 9, we call this the nested HDP (nHDP), and write $\theta'_{ji} \sim nHDP(\{\alpha'_j\}, \gamma', \{\alpha_{k'}\}, \gamma, H)$. Similar to the nDP, and the DP-HDP (Eqn. 11), this again defines a distribution over the space of distributions over distributions. The complete nHDP mixture model is defined by subsequently sampling $\theta_{ji} \sim \theta'_{ji}$, followed by $x_{ji} \sim F(\theta_{ji})$.

An alternative characterization of the nHDP mixture model is using the topic index z_{ji} and entity index z'_{ji} corresponding to x_{ji}:

$$\begin{aligned} \beta \sim GEM(\gamma);\ \pi_{k'} \sim DP(\alpha, \beta);\ \phi_k \sim H,\ k, k' = 1 \ldots \infty \\ \beta' \sim GEM(\gamma')\ ; \pi'_j \sim DP(\alpha, \beta'),\ j = 1 \ldots M \\ z'_{ji} \sim \pi'_j\ ;\ z_{ji} \sim \pi_{z'_{ji}};\ x_{ji} \sim F(\phi_{z_{ji}}),\ i = 1 \ldots n_j \end{aligned} \tag{14}$$

This may be understood as first creating entity-specific distributions $\pi_{k'}$ over topics using global topic popularities β, followed by creation of document-specific distributions π'_j over entities using global entity popularities β'. Using these parameters, the content of the j^{th} document is generated by sampling repeatedly in iid fashion an entity index z'_{ji} using π'_j, a topic index z_{ji} using $\pi_{z'_{ji}}$ and finally a word using $F(\phi_{z_{ji}})$.

Observe the connection with the ATM in Eqn. 1. The main difference is the the set of entities and topics is infinite. Additionally, each document now has a distinct non-uniform distribution π'_j over entities.

Also, observe that we have preserved the HDP notation to the extent possible, to facilitate understanding. To distinguish between variables corresponding to the two HDPs in the model, we use dashes ($'$) as superscripts on symbols corresponding to the outer HDP. Going forward, we follow the same convention for naming variables in the nested CRF.

Nested Chinese Restaurant Franchise: In this section, we derive the predictive distribution for the next draw θ'_{ji} from the nHDP given previous draws, after integrating out $\{G'_j\}$ and G'_0, and then the predictive distribution for the draw θ_{ji} after integrating out $\{\theta'_{ji}\}$ and G_0. We also provide an interpretation for these using two nested CRFs, corresponding to the inner and outer HDPs. These will be useful for the inference algorithm that we describe in Section 5.

Let $\{\theta'_{ji}\}$ denote the sequence of draws from G'_j, and $\{\psi'_{jt'}\}_{t'=1}$ denote the sequence of draws from G'_0. Then the conditional distribution of θ'_{ji} given all previous draws after integrating out G'_j looks as follows:

$$\theta'_{ji}|\theta'_{j1:i-1}, \alpha'_j, G'_0 \sim \sum_{t'=1}^{m'_{j\cdot}} \frac{n'_{jt'\cdot}}{i-1+\alpha'_j} \delta_{\psi'_{jt'}} + \frac{\alpha'_j}{i-1+\alpha'_j} G'_0 \tag{15}$$

where $n'_{jt'k'} = \sum_{i'} \delta(\theta'_{ji'}, \psi'_{jt'})\delta(\psi'_{jt'}, G_{k'})$, $m'_{jk'} = \sum_{t'} \delta(\psi'_{jt'}, G_{k'})$. Next, we integrate out G'_0, which is also distributed according to Dirichlet process:

$$\begin{aligned} \psi'_{jt'}|\psi'_{11}, \psi'_{12}, \ldots, \psi'_{21}, \ldots, \psi'_{j,t'-1}, \gamma', \mathrm{HDP}(\boldsymbol{\alpha}, \gamma, H) \sim \\ \sum_{k'=1}^{K'} \frac{m'_{\cdot j}}{m'_{\cdot\cdot} + \gamma'} \delta_{G_{k'}} + \frac{\gamma'}{m'_{\cdot\cdot} + \gamma'} \mathrm{HDP}(\boldsymbol{\alpha}, \gamma, H) \end{aligned} \tag{16}$$

Observe that each θ'_{ji} variable gets assigned to one of the $G_{k'}$ variables. Let $\{\theta_{ji}\}$ denote the sequence of draws from respective $\{\theta'_{ji}\}$ (i.e. from the corresponding $G_{k'}$), $\{\psi_{k't}\}$ the sequence of draws from G_0, and $\{\phi_k\}_{k=1}^{\infty}$ the sequence of draws from H. Let $\theta_{k':ji}$ denote the set of θ variables already drawn from $G_{k'}$ before sampling θ_{ji}, i.e. $\theta_{k':ji} \equiv \{\theta_{j'i'} : \theta'_{j'i'} = G'_k, \ \forall i', \ j' \leq j, \text{ and } i' < i, \ j' = j\}$. Then, the conditional distribution of θ_{ji} given $\theta_{k':ji}$ and G_0, after integrating out $G_{k'}$ (corresponding to θ'_{ji}) is as follows:

$$\theta_{ji}|\theta_{k':ji}, \alpha_0, G_0 \sim \sum_{t=1}^{m_{k'\cdot}} \frac{n_{k't\cdot}}{i-1+\alpha_0} \delta_{\psi_{k't}} + \frac{\alpha_0}{n_{k'\cdot\cdot} + \alpha_0} G_0 \tag{17}$$

where $n_{k'tk} = \sum_i \delta(\theta_{k'i,\psi_{k't}})\delta(\psi_{k't}, \phi_k)$, $m_{k'k} = \sum_t \delta(\psi_{k't}, \phi_k)$ and dots indicate marginal counts. As G_0 is also distributed according to a Dirichlet Process, we can integrate it out similarly and write the conditional distribution of $\psi_{k't}$ as follows:

$$\psi_{k't}|\psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{k't-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma} H \tag{18}$$

Note that both conditional distributions for θ'_{ji} and θ_{ji} are similar to that for CRF (Eqns. 6 and 7). We interpret these two distributions as a *nested Chinese Restaurant Franchise*, involving one inner CRF and one outer CRF.

Consider a set of outer restaurants, one corresponding to each group. Customers entering each of these restaurants select a table θ'_{ji} according a group specific CRP (Eqn 15). The restaurants share a common set of inner restaurants $\{G_{k'}\}$. Inner restaurants are assigned to the tables of each outer restaurant according to another CRP (Eqn 16). Next, the customers go to the inner restaurant assigned to them (by some outer restaurant) and select a table θ_{ji} according to the inner restaurant specific CRP (Eqn 17). These inner restaurants share a common menu of dishes $\{\phi_k\}$. Dishes are assigned to the tables of each inner restaurant according to another CRP (Eqn 18).

Let t'_{ji} be the (outer table) index of the $\psi'_{jt'}$ associated with θ'_{ji}, and let k'_{jt} be the (inner restaurant) index of the $G_{k'}$ associated with $\psi'_{jt'}$. Let t_{ji} be the (inner table) index of the $\psi_{k't}$ associated with θ_{ji}, and let $k_{k't}$ be the (dish) index of the ϕ_k associated with $\psi_{k't}$. Then the two conditional distributions above can also be written in terms of the indexes $\{t'_{ji}\}, \{k'_{jt'}\}$, $\{t_{ji}\}$ and $\{k_{k't}\}$ instead of referring to the distributions directly. For the j^{th} outer restaurant and its i^{th} customer, we draw θ'_{ji} using Eqn. 15. If the first summation is chosen, we set $\theta'_{ji} = \psi'_{jt'}$ and let $t'_{ji} = t'$ for the chosen t'. If the second term is chosen, then we increment $m'_{j\cdot}$ by one, draw $\psi'_{jm'_{j\cdot}} \sim G'_0$ using (Eqn 16) and set $\theta'_{ji} = \psi'_{jm'_{j\cdot}}$ and $t'_{ji} = m'_{j\cdot}$. If we draw $\psi'_{jt'}$ via choosing a summation term, we set $\psi'_{jt'} = G_{k'}$ and let $k'_{jt'} = k'$ for the chosen k'. If the second term is chosen, then we increment the current distinct entity count M by one, draw $G_M \sim \text{HDP}(\boldsymbol{\alpha}, \gamma, H)$ and set $\psi'_{jt'} = G_M$ and $k'_{jt'} = M$. Next, we similarly draw samples of θ_{ji} for each j and i using Eqn. 17. If new sample from G_0 is needed, we use Eqn. 18 to obtain a new sample $\psi_{k't}$.

5 Inference for Nested HDP

We use Gibbs sampling for approximate inference, as exact inference is intractable for this problem. The conditional distributions for the nCRF scheme lend themselves to an inference algorithm where we sample t'_{ji}, t_{ji}, $k'_{jt'}$ and $k_{k't}$. The conditionals for these variables are similar to those in equations 15, 17, 18 and 16 respectively. However, in such an approach, there exists a tight coupling between the variables t'_{ji}, t_{ji} and $k'_{jt'}$, which would call for computationally expensive joint sampling of variables.

Instead, we adopt a technique similar to the direct sampling scheme in HDP[16], where variables G_0 and G'_0 are explicitly sampled instead of being integrated out, by sampling the stick breaking weights β and β' respectively. Further, we directly sample z_{ji} (the topic) and z'_{ji} (the author) for each word in the j^{th} document avoiding explicit table assignments to the t'_{ji} and t_{ji} variables. However, in order to sample β and β', the table information is maintained in the form of the number of tables in each outer and inner restaurant, $m'_{jk'}$ and $m_{k'k}$ respectively. Thus the latent variables that need to be sampled in our Gibbs sampling scheme are z_{ji}, z'_{ji}, β, β', $m_{jk'}$ and $m_{k'k}$.

We introduce the following notation for the rest of this section. Let $\mathbf{x} = \{x_{ji} : \text{all } j, i\}$, $\mathbf{x_{-ji}} = \{x_{j'i'} : j' \neq j, i' \neq i\}$, $\mathbf{m} = \{m_{k'k} : \text{all } k', k\}$, $\mathbf{m'} = \{m'_{jk'} : \text{all } j, k'\}$, $\mathbf{z} = \{z_{ji} : \text{all } j, i\}$, $\mathbf{z'} = \{z'_{ji} : \text{all } j, i\}$, $\mathbf{z_{-ji}} = \{z_{j'i'} : j' \neq j, i' \neq i\}$, $\mathbf{z'_{-ji}} = \{z'_{j'i'} : j' \neq j, i' \neq i\}$, $\beta_{new} = \sum_{k=K+1}^{\infty} \beta_k$ and $\beta'_{new} = \sum_{k'=K'+1}^{\infty} \beta'_{k'}$

Sampling z_{ji}: The conditional distribution for topic index z_{ji} is

$$p(z_{ji} = z | \mathbf{z_{-ji}}, \mathbf{z'_{-ji}}, z'_{ji} = z', \mathbf{m}, \mathbf{m'}, \beta, \beta', \mathbf{x}) \propto p(z_{ji} = z | \mathbf{z_{-ji}}, z'_{ji} = z') p(x_{ji} | z_{ji} = z, \mathbf{x_{-ji}})$$

For existing topics, $p(z_{ji} = z | \mathbf{z_{-ji}}, z'_{ji} = z')$ can be split into two terms, one from picking any of the existing tables from entity (inner restaurant) z' with topic z and the other from creating a new table for entity z' and assigning the topic z to it. For a new topic, a new table is always created for entity z'. Hence,

$$p(z_{ji} = z|\mathbf{z_{-ji}}, z'_{ji} = z') \propto \begin{cases} \frac{n_{k'.k}+\alpha\beta_k}{n_{k'.}+\alpha} & \text{Existing } z \\ \frac{\alpha\beta_{new}}{n_{k'..}+\alpha} & \text{New topic } z \end{cases} \tag{19}$$

The other term $p(x_{ji}|z_{ji} = z, \mathbf{x_{-ji}})$ is the conditional density of x_{ji} under topic z given all data items except x_{ji}. Assuming each topic is sampled from a V dimensional symmetric Dirichlet prior over the vocabulary with parameter η, i.e $\phi_k \sim Dir(\eta)$, the above probability can be simplified to the following expression, by integrating out ϕ:

$$p(x_{ji} = w|z_{ji} = z, \mathbf{x_{-ji}}) \propto \frac{n_{zw} + \eta}{n_{z\cdot} + V\eta}$$

where n_{zw} is the number of occurrences of topic z with word w in the vocabulary.

Sampling z'_{ji}: The conditional distribution for the entity index z'_{ji} is

$$p(z'_{ji} = z'|\mathbf{z'_{-ji}}, \mathbf{z_{-ji}}, z_{ji} = z, \mathbf{m}, \mathbf{m'}, \beta, \beta', \mathbf{x}) \propto p(z'_{ji} = z|\mathbf{z'_{-ji}})p(z_{ji} = z|\mathbf{z_{-ji}}, z'_{ji} = z')$$

Again, $p(z_{ji} = z|\mathbf{z_{-ji}}, z'_{ji} = z')$ can be split into two terms, one from picking an existing outer table with entity z' and the other from creating a new outer table and assigning the entity z' to it. Further, creation of a new entity always involves the creation of a new outer table. Hence,

$$p(z'_{ji} = z'|\mathbf{z'_{-ji}}) \propto \begin{cases} \frac{n'_{j.k'}+\alpha'\beta'_{k'}}{n'_{j..}+\alpha'} & \text{Existing } z' \\ \frac{\alpha'\beta'_{new}}{n'_{j..}+\alpha'} & \text{New } z' \end{cases}$$

$p(z_{ji} = z|\mathbf{z_{-ji}}, z'_{ji} = z')$ follows from Eqn. 19.

Sampling β and β': The posterior of G_0, conditioned on samples $\psi_{k't}$ from it, is also distributed as a DP due to Dirichlet-Multinomial conjugacy, and the stick breaking weights of G_0 can be sampled as follows: $(\beta_1, \beta_2 \ldots \beta_K, \beta_{new}) \sim Dir(m_{.1}, m_{.2} \ldots m_{.K}, \gamma)$ Similarly, the stick breaking weights β' can be sampled from the posterior distribution of G'_0 conditioned on samples from G'_0 in the form of $m'_{jk'}$ as follows: $(\beta'_1, \beta'_2 \ldots \beta'_K, \beta'_{new}) \sim Dir(m'_{.1}, m'_{.2} \ldots m'_{.K'}, \gamma')$

Sampling m and m': $m_{k'k}$ is the number of inner tables generated as $n_{k'.k}$ samples are drawn from G'_k corresponding to a particular topic k. This is the number of partitions generated as samples are drawn from a Dirichlet Process with concentration parameter $\alpha\beta_k$ and is distributed according to a closed form expression [1]. We adopt a different method [7] for sampling $m_{k'k}$ by drawing a total of $n_{k'.k}$ samples with topic k, and incrementing the count $m_{k'k}$ whenever a new inner table is created with topic assignment k. Similarly, $m'_{jk'}$ is sampled by drawing a total of $n'_{j.k'}$ samples with entity k' and incrementing the count $m'_{jk'}$ whenever a new outer table is created with entity assignment k'.

Sampling Concentration parameters: We place a vague gamma prior on the concentration parameters $\alpha, \gamma, \alpha', \gamma'$ with hyper parameters (α_a, α_b), (γ_a, γ_b), (α'_a, α'_b) and (γ'_a, γ'_b) respectively. We use Gibbs sampling scheme for sampling the concentration parameters using the technique outlined in HDP[16].

We use the conditional distributions above to perform inference under three different settings. In the **"no observed entities"** setting, the conditional distributions above are repeatedly sampled from until convergence. For initialization, we first initialize the topic variables z_{ji} using an online scheme, and then initialize the entities z'_{ji} using the topics. In the **"completely observed entities"** setting, the set of entities A_j is given for every document j. Since no other entities are deemed possible for the j^{th} document, $p(z'_{ji} = z' | \mathbf{z'_{-ji}}, \mathbf{z_{-ji}}, z_{ji} = z, \mathbf{m}, \mathbf{m'}, \beta, \beta', \mathbf{x})$ is set to 0 for new entities and for all $z' \notin A_j$. In the **"partially observed entities"** setting, a partial list of known entities A_j is available for document j, but other entities are also considered possible. We perform an initialization step, similar to that in the completely observed setting, using the known entities A_j. No new authors are added in this initial step. During later iterations, we allow all assignments $z'_{ji} = z$ — one of the known entities from A_j, entities of other documents j', and new entities. However, we introduce a bias towards the known authors $z' \in A_j$ using an additional small positive term to their probability mass.

6 Experiments

In this section, we experimentally evaluate the proposed nHDP model for the task of modeling author entities who have collaboratively written research papers, and compare its performance against available baselines. Specifically, we evaluate two different aspects: (1) how well the model is able to learn from the training samples and fit held-out data, first (1a) when all the authors are observed in training and test documents, and secondly (1b) when some or all of the authors are unobserved in training and test documents, (2) how accurately the model discovers hidden authors, who are not mentioned at all in the corpus.

We consider the following models for the experiments: (i) The author-topic model (**ATM**) (Eqn. 1) where the number of topics is pre-specified, and all authors are observed for all documents. This is used as a baseline for (1a) above. (ii) The Hierarchical Dirichlet Process (**HDP**) (Eqn. 4) using the direct assignment inference scheme for fair comparison. We use our own implementation for this. Recall that the HDP infers the number of topics, and does not use author information. (iii) nHDP with completely observed entities (**nHDP-co**), which assumes complete entity information to be available for all documents, but learns topics in a nonparametric fashion. This can be imagined as an improvement over ATM where the number of topics does not need to be specified. (iv) nHDP with partially observed entities (**nHDP-po**), which makes use of available entity information, but admits the possibility of entities being hidden globally from the corpus, or locally from individual documents. (v) nHDP with no observed entities (**nHDP-no**), which does not make use of any entity information and assumes all entities to be globally hidden in the corpus. For task (1a) above, the applicable models are the ATM, HDP (which ignores the entity information) and nHDP-co. For task (1b), the ATM does not apply. We evaluate

Table 1. Perplexity of ATM, HDP and nHDP-co for NIPS

Model	ATM	HDP	nHDP-co
Perplexity	2783	1775	1247

HDP, nHDP-po and nHDP-no. It is important to point out that there are no available baselines for task (2) above.

We use the following publicly available publication datasets for our experimental analysis. The **NIPS** dataset[3] is a collection of papers from Neural Information Processing Systems (NIPS) conference proceedings (volume 0-12). This collection contains 1,740 documents contributed by a total of 2,037 authors, with total 2,301,375 word tokens resulting in a vocabulary of 13,649 words. A subset of the **DBLP Abstracts** dataset[4] containing 12,000 documents by 15,252 authors collected from 20 conferences records on the Digital Bibliography and Library Project (DBLP). Each document is represented as a bag of words present in abstract and title of the corresponding paper, resulting in a vocabulary size of 11,771 words.

1. Generalization Ability: We now come to our first experiment, where we evaluate the ability of the models, whose parameters are learnt from a training set, to predict words in new unseen documents in a held-out test set. We evaluate performance of a model M on a test collection D using the standard notion of perplexity [3]: $exp(-\sum_{d \in D} p(w_d)|M)$.

In experiment (1a), all authors are observed in training and test documents. To favor the ATM, which cannot handle new authors in test document, we create test-train splits ensuring that each author in the test collection occurs in at least one training document.

Perplexity results are shown in Table 1. Recall that HDP and nHDP find the best number of topics, while for ATM we have recorded its best performance across different value of K. The results show that while knowledge of authors is useful, the ability of non-parametric topic models to infer the number of topics clearly leads to better generalization.

Next, in experiment (1b), we first create training-test distributions with reasonable author overlap by letting each author vote with probability 0.7 whether to send a document to train or test, and majority decision is taken for each document. Next, authors are partially hidden from both the test and the train documents as follows. We iterate over the global list of authors and remove each author from all training and test documents with probability p_g. We then iterate over each training and test document, and remove each remaining author of that document with probability p_l. We experiment with different values of p_g and p_l to simulate different extents of missing information on authors.

The results are shown in Table 2. We can see that when more information is available about the authors, the ability to fit held-out data improves. More

[3] http://www.arbylon.net/resources.html

[4] http://www.cs.uiuc.edu/ hbdeng/data/kdd2011.htm

Table 2. Perplexity for HDP and nHDP with varying percentage of hidden authors

Model	HDP	nHDP-no	nHDP-po	nHDP-po	nHDP-po	nHDP-co
p_g,p_l	1,1	1,1	0.6,0.6	0.4,0.4	0.2,0.2	0,0
Perplexity NIPS	2572	1882	1434	1266	1109	987
Perplexity DBLP	1027	997	935	869	676	394

interestingly, even when no / very little author information is available, just the assumption about the existence of authors, or a discrete set of topic mixtures, leads to better generalization ability, as can be seen from the relative performance of HDP and nHDP-no.

2. Discovering Missing Authors: Beyond data fitting, the most significant ability of the nHDP mixture model is to discover entities which are relevant for documents in the corpus, but are never mentioned. We perform a case study with the top 6 most prolific authors in NIPS, by removing them completely from the corpus, and then checking the ability of the model to discover them in a completely unsupervised fashion. While it is possible to define as a classification problem the task of detecting of *locally missing* authors in individual documents when the author is observed in other documents, we reiterate that there is no existing baseline when an author is *globally hidden*.

We evaluate the accuracy of discovering hidden author as follows. For each hidden author $h \in \{1 \ldots H\}$, we create a m-dimensional vector c_h, where m is the corpus size, with $c_h[j]$ indicating his authorship in the j^{th} document. We explored two possibilities for this 'true' indicator vector: (a) binary indicators using the gold-standard author names for documents, and (b) the number of words written by that author in the document according to nHDP with completely observed authors (nHDP-co). Similarly, we create an m-dimensional vector for each new author $n \in \{1 \ldots N\}$ discovered by the nHDP-po, with $c_n[j]$ indicating his contribution (no. of authored words) in the j^{th} document. We now check how well the vectors $\{c_n\}$ correspond to the 'true' vectors $\{c_h\}$. This is done by defining two variables C_n and C_h, taking values $1 \ldots H$ and $1 \ldots N$ respectively, and defining a joint distribution over them as $P(h,n) = \frac{1}{Z}\text{sim}(c_h, c_n)$, where Z is a normalization constant. For $\text{sim}(c_h, c_n)$, we use cosine similarity between normalized versions of c_h and c_n. Mutual information $I(C_h, C_n) = \sum_{h,n} p(h,n) \log \frac{p(h,n)}{p(h)p(n)}$ measures the information that C_h and C_y share. We used its normalized variant $NMI(C_h, C_n) = \frac{I(C_h,C_n)}{|H(C_h)+H(C_n)|/2}$ ($H(X)$ indicating entropy of X) which takes values between 0 and 1, higher values indicating more shared information.

First, we note that the best NMI achievable for this task, by replacing the true vectors $\{c_h\}$ for the discovered vectors $\{c_n\}$, is 0.86 for case (a) and 0.98 for case (b) above. In comparison, using nHDP-po, we achieve NMI scores of 0.59 for case (a) and 0.72 for case (b). This indicates that the actual author distributions that the model discovers not only help in fitting the data, but also have reasonable correspondence with the true hidden authors. We believe that this is a promising initial step in addressing this difficult problem.

7 Conclusions

In this paper, we have addressed the problem of entity-topic analysis from document corpora, where the set of document entities are either completely or partially hidden. For such problems, we have proposed as a prior distribution the nested Hierarchical Dirichlet Process, which consists of two levels of Hierarchical Dirichlet Processes, where one is the base distribution of the other. Using a direct sampling scheme for inference, we have shown that the nHDP is able to generalize better than existing models under varying available knowledge about authors in research publications, and is additionally able to discover completely hidden authors in the corpus.

References

1. C. Antoniak. Mixtures of Dirichlet Processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 2(6):1152–1174, 1974.
2. D. Blei, T. Griffiths, M. Jordan, and J. Tanenbaum. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *JACM*, 2010.
3. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 2003.
4. A. Dai and A. Storkey. Author disambiguation: A nonparametric topic and co-authorship model. In *NIPS Workshop on Applications for Topic Models Text and Beyond, pages 1–4*, 2009.
5. E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *PNAS*, 101(Suppl 1), 2004.
6. T. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230, 1973.
7. E. Fox, E. Sudderth, M. Jordan, and A. Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Stats.*, 5(2A):1020–1056, 2011.
8. H. Kim, Y. Sun, J. Hockenmaier, and J. Han. Etm: Entity topic models for mining documents associated with entities. In *ICDM*, pages 349–358, 2012.
9. A. McCallum, A. Corrada-Emmanuel, and X. Wang. The author recepient topic model for topic and role discovery in social networks. 2004.
10. D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *ACM SIGKDD*, KDD '06, pages 680–686, New York, NY, USA, 2006. ACM.
11. J. Paisley, C. Wang, D. Blei, and M. Jordan. Nested hierarchical dirichlet processes. *Arxiv*, 2012.
12. J. Pitman. Gibbs sampling methods for stick-breaking priors. *Lecture Notes for St. Flour Summer School*, 2002.
13. A. Rodriguez, D. Dunson, and A. Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.
14. M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.
15. J. Sethuraman. A constructive definition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.
16. Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006.
17. D. Wulsin, S. Jensen, and B. Litt. A hierarchical dirichlet process model with multiple levels of clustering for human eeg seizure modeling. In *ICML*, 2012.