Dynamic Hierarchical Clustering for Data Exploration

Avinava Dubey
Machine Learning Department
Carnegie Mellon University
akdubey@cs.cmu.edu

Abstract

Background Hierarchical clustering methods offer an intuitive and powerful way to model and explore a wide variety of data sets. However, the assumption of a fixed hierarchy is often overly restrictive as the hierarchy may evolve when working with data generated over a period of time. Due to this restriction, existing methods that can only learn static hierarchies and are unable to model hierarchies that evolve over time.

Aim In this paper we aim to overcome this problem. We expect both the structure of our hierarchy and the parameters of the clusters to evolve with time. To this end we aim to model this evolution of both structure of hierarchy and the parameters over time. Our primary goal in this paper is to show that doing so will lead to better modeling of such data and help us in discovering interesting trends in data.

Data We explore three different datasets: (a) 79,800 paper titles from the Proceedings of the National Academy of Sciences (PNAS) between 1915 and 2005, containing 36,901 unique words (b) Presidential State of the Union (SoU) addresses from 1790 through 2002, containing 56,352 sentences and 21,505 unique words (c) 673,102 tweets containing hashtags relevant to the NFL, collected over 18 weeks in 2011 and containing 2,636 unique words.

Methods In this paper, we define a distribution over temporally varying trees with infinitely many nodes (representing clusters) that captures evolving hierarchies. We show that this model can be used to cluster both real-valued and discrete observations. Finally, we propose a scalable approximate Markov chain Monte Carlo inference scheme that can be run in a distributed manner.

Results Quantitatively, we find that our model has significantly better log-likelihood on test data for all three datasets considered. Qualitatively, we also found interesting sub-parts of the hierarchies and how they evolve over time. For example, we show how "Immunology" appeared and then evolved over time in the PNAS dataset after the discovery of the structure of antibodies in 1960.

Conclusion The quantitative experiment shows that as compared to baselines our dynamic hierarchical clustering better fits a given dataset and generalizes well to unseen data. We also show that we are able to find interesting trends that have not been seen before in the given datasets. Through our experiments we make a compelling case for using dynamic hierarchical clustering models to explore data generated over a period of time.

1 Introduction

Hierarchically structured clustering models offer a natural representation for many forms of data. For example, we may wish to cluster animals in a hierarchical manner, where for example "dog" and "cat" are subcategories of "mammal", and "poodle" and "dachshund" are subcategories of "dog". When modeling scientific articles, articles about machine learning and articles about programming languages may be categorized in subcategories of articles about computer science. Representing clusters in a tree structure allows us to explicitly capture these relationships, and allow clusters that are closer in tree-distance to have more similar parameters.

Since such hierarchical structures occur commonly in the real world, it is not surprising that there exists a rich literature on statistical models for trees. We are particularly interested in *nonparametric* distributions over trees – that is, distributions over trees with infinitely many leaves and infinitely many internal nodes. We can model any finite data set using a finite subset of such a tree, marginalizing over the infinitely many unoccupied branches. The advantage of such an approach is that we do not have to specify the tree dimensionality in advance, and can grow our representation in a consistent manner if we observe more data.

In many settings, our data points are associated with a point in time – for example the date when a photograph was taken or an article was written. A stationary clustering model is inappropriate in such a context: The number of clusters may change over time; the relative popularities of clusters may vary; and the location of each cluster in parameter space may change. As an example, consider a topic model for scientific articles over the twentieth century. The field of computer science – and therefore topics related to it – did not exist in the first half of the century. The proportion of scientific articles devoted to genetics has likely increased over the century, and the terminology used in such articles has changed with the development of new sequencing technology.

Despite this, to the best of our knowledge, there are no nonparametric distributions over time-evolving trees in the literature. There exist a variety of distributions over *stationary* trees [Adams et al., 2010, Rodriguez et al., 2008, Blei et al., 2004, Neal, 2003, Kingman, 1982], and time-evolving non-hierarchical clustering models [Wang and McCallum, 2006, Caron et al., 2007, Lin et al., 2010, Ahmed and Xing, 2008, Blei and Frazier, 2011, MacEachern, 1999, Dubey et al., 2013] – but no models that combine time evolution and hierarchical structure. The reason for this is likely to be practical: Inference in trees is typically very computationally intensive, and adding temporal variation will, in general, increase the computational requirements. Designing such a model must, therefore, proceed hand in hand with developing efficient and scalable inference schemes.

In this paper, we define a distribution over temporally varying trees with infinitely many nodes that captures the form of variation above. We describe how this model can be used to cluster both real-valued observations and text data. Further, we propose a scalable approximate inference scheme that can be run in a distributed manner. We demonstrate the efficacy of this inference scheme on synthetic data where ground-truth clustering is available, and demonstrate qualitative and quantitative performance on three text corpora.

2 Background

A *dependent Dirichlet process* [MacEachern, 1999] is a distribution over collections of probability measures, that vary with time or some other covariate, and has the property that, at any time point, the marginal distribution over the probability measure at that time point is given by a Dirichlet process [Ferguson, 1973]. This idea can be extended to other nonparametric processes: A dependent Pitman-Yor process [Sudderth and Jordan, 2008] is a distribution over collections of probability measures whose marginals are given by the Pitman-Yor process [Pitman and Yor, 1997]; a dependent Indian buffet process [Williamson et al., 2010] is a distribution over collections of binary matrices whose marginals are given by the Indian buffet process. There exist many similarities in the constructions of such dependent nonparametric processes; in Section 2.2 we discuss some distributions for nonparametric time-varying clustering models that share properties with our model.

The key difference between our proposed model and the existing range of dependent nonparametric models is that our model has tree-distributed marginals. There exist a number of choices for the marginal distribution over trees, as we discuss in Section 2.2. We have chosen to use a distribution over infinite-dimensional trees known as the Tree Structured Stick Breaking Process `TSSBP` [Adams et al., 2010], which we describe in Section 2.1. Unlike other distributions over infinite-dimensional trees, the TSSBP allows data to be associated with internal nodes as well as leaves. We discuss how this may be desirable in Section 2.1.

2.1 The tree-structured stick-breaking process

The tree-structured stick-breaking process (TSSBP) is a distribution over trees with infinitely many leaves and infinitely many nodes. Each node within the tree is associated with a mass π_ϵ such that $\sum_\epsilon \pi_\epsilon = 1$, and each data point is assigned to a node in the tree according to

$$p(z_n = \epsilon) = \pi_\epsilon,$$

where z_n is the node assignment of the nth data point. The TSSBP is unique among the current toolbox of random infinite-dimensional trees, in that data can be assigned to an internal node, rather than a leaf, of the tree. This property is often desirable: For example in a topic modeling context, a document could be assigned to a general topic such as "science" that lives toward the root of the tree, or to a more specific topic such as "genetics" that is a descendant of the science topic.

The TSSBP can be represented using two interleaving stick-breaking processes – one (parametrized by α) that determines the size of a node and another (parametrized by γ) that determines the branching probabilities. Index the root node as node $\emptyset$ and let $\pi_\emptyset$ be the mass assigned to it. Index its (countably infinite) child nodes as node 1, node $2, \ldots$ and let $\pi_1, \pi_2, \ldots$ be the masses assigned to them; index the child nodes of node 1 as nodes $1 \cdot 1, 1 \cdot 2, \ldots$ and let $\pi_{1\cdot 1}, \pi_{1\cdot 2}, \ldots$ be the masses assigned to nodes $1 \cdot 1, 1 \cdot 2 \ldots$; etc. Then

we can sample the infinite-dimensional tree as:

$$\begin{aligned}
\nu_\epsilon &\sim \text{Beta}(1, \alpha(|\epsilon|)) \\
\psi_\epsilon &\sim \text{Beta}(1, \gamma) \\
\pi_\emptyset &= \nu_\emptyset, \quad \phi_\emptyset = 1 \\
\phi_{\epsilon \cdot i} &= \psi_{\epsilon \cdot i} \textstyle\prod_{j=1}^{i-1}(1 - \psi_{\epsilon \cdot j}) \\
\pi_\epsilon &= \nu_\epsilon \phi_\epsilon \textstyle\prod_{\epsilon' \prec \epsilon}(1 - \nu_{\epsilon'})\phi_{\epsilon'},
\end{aligned} \tag{1}$$

where $|\epsilon|$ indicates the depth of node ϵ, and $\epsilon' \prec \epsilon$ indicates that ϵ' is an ancestor node of ϵ. We refer to the resulting infinite-dimensional weighted tree as $\Pi = ((\pi_\epsilon), (\phi_{\epsilon i}))$.

The resulting weighted tree-structure can be used to create distributions over hierarchically arranged parameters, which can in turn can be used to create distributions over hierarchically arranged clusters. Unlike other nonparametric tree structures, clusters are not constrained to lie at leaf nodes. In a topic modeling context, this means that a document could be assigned to a general topic such as "science" that lives toward the root of the tree, or to a more specific topic such as "genetics" that is a descendant of the science topic.

2.2 Other related work

In this paper, we propose a method for a temporally varying, tree-structured clustering model with an unbounded number of clusters. A number of existing models incorporate one or more of these features.

There exist a wide variety of distributions over trees with infinitely many nodes, including the nested Chinese restaurant process [Blei et al., 2004], the Dirichlet diffusion tree [Neal, 2003], and Kingman's coalescent [Kingman, 1982]. These models differ from the TSSBP in that data can only be associated with a leaf node, or equivalently a full path from root to leaf. We chose to base our clustering model on the TSSBP because, in many applications, it makes sense to associate data with internal nodes. For example, a document may be narrowly about Physics or Biology, or may be a more broad article on the sciences in general.

While, to the best of our knowledge, there exist no temporally varying nonparametric tree distributions, there do exist a wide variety of temporally varying nonparametric clustering models, several of which are related to the model proposed in this paper. The dependent Dirichlet process models of [Caron et al., 2007] and [Lin et al., 2010] specify a distribution over clusterings of data, where the popularity of a cluster can vary over time. These models are based on the Chinese restaurant process: the probability of joining a cluster at time t depends on both the number of words associated with that topic at time t (as in the standard Chinese restaurant process), *and* on the word counts from previous time periods. We modify this approach to allow the node weights in our sequence of trees to vary over time.

Other models have been used to allow the parameters associated with clusters to vary over time. The single-p dependent Dirichlet process [MacEachern, 1999] clusters data according to a Dirichlet process, and evolves the cluster parameters according to a stochastic process. In a parametric setting that is similar to the topic model proposed in Section 3.3, the Dynamic Topic Model [Blei and Lafferty, 2006], parametrizes each topic, or cluster, using a logistic normal distribution. Time dependence is induced by allowing the underlying Gaussian-distributed vector to evolve via multivariate increments.

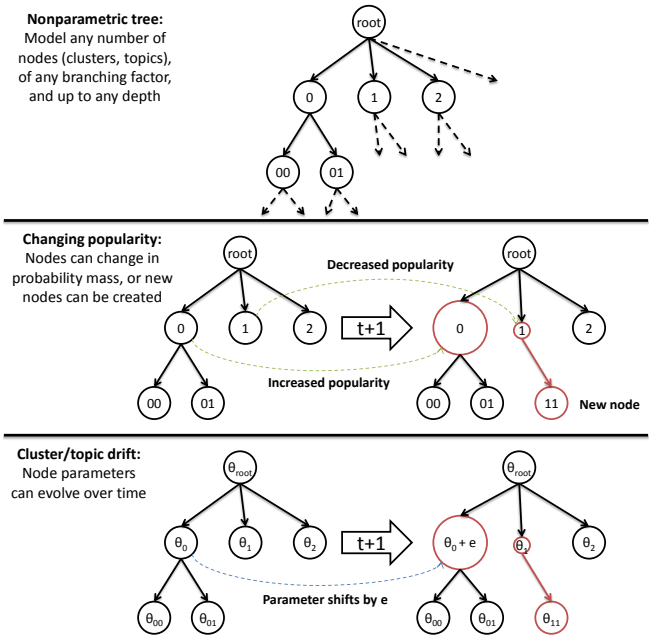

Figure 1: Our dependent tree-structured stick breaking process can model trees of arbitrary size and shape, and captures popularity and parameter changes through time.

3 Dependent tree-structured stick-breaking processes

A dependent nonparametric process [MacEachern, 1999] is a distribution over collections of random measures indexed by values in some covariate space, such that at each covariate value, the marginal distribution is given by some known nonparametric distribution. If our covariate space is time, this results in a distribution over time-varying random measures. There are two main methods of inducing dependency: Allowing the sizes of the atoms composing the measure to vary across covariate space, and allowing the parameter values associated with the atoms to vary across covariate space. In the context of a time-dependent topic model, these methods correspond to allowing the popularity of a topic to change over time, and allowing the words used to express a topic to change over time (topic drift).

In this section, we describe a dependent tree-structured stick-breaking process where both the atom sizes and their locations vary with time. We begin by describing the distribution over atom sizes, and then use this distribution over collections of trees as the basis for temporally varying clustering models and topic models.

3.1 A distribution over time-varying trees

We start with the basic TSSBP model of [Adams et al., 2010], described in Section 2.1 and represented in the top row of Figure 1, and modify it so that the latent variables ν_ϵ, ψ_ϵ and π_ϵ are replaced with sequences $\nu_\epsilon^{(t)}$, $\psi_\epsilon^{(t)}$ and $\pi_\epsilon^{(t)}$ indexed by discrete time $t \in \mathcal{T}$ (as represented in the middle row of Figure 1). The forms of $\nu_\epsilon^{(t)}$ and $\psi_\epsilon^{(t)}$ are chosen so that the marginal distribution over the $\pi_\epsilon^{(t)}$ is as described in Equation 1.

Let $N^{(t)}$ be the number of observations at time t, and let $z_n^{(t)}$ be the node allocation of the nth observation at time t. For each node ϵ at time t, let $X_\epsilon^{(t)} = \sum_{n=1}^{N_t} \mathbb{I}(z_n^{(t)} = \epsilon)$ be the number of observations assigned to node ϵ at time t, and $Y_\epsilon^{(t)} = \sum_{n=1}^{N_t} \mathbb{I}(\epsilon \prec z_n^{(t)})$ be the number of observations assigned to descendants of node ϵ. Introduce a "window" parameter $h \in \mathbb{N}$. We can then define a prior predictive distribution over the tree at time t, as

$$\begin{aligned} \nu_\epsilon^{(t)} &\sim \text{Beta}\big(1 + \textstyle\sum_{t'=t-h}^{t-1} X_\epsilon^{(t')}, \alpha(|\epsilon|) + \sum_{t'=t-h}^{t-1} Y_\epsilon^{(t')}\big) \\ \psi_{\epsilon\cdot i}^{(t)} &\sim \text{Beta}\big(1 + \textstyle\sum_{t'=t-h}^{t-1}(X_{\epsilon\cdot i}^{(t')} + Y_{\epsilon\cdot i}^{(t')}), \\ &\qquad \gamma + \textstyle\sum_{j>i} \sum_{t'=t-h}^{t}(X_{\epsilon\cdot j}^{(t')} + Y_{\epsilon\cdot j}^{(t')})\big). \end{aligned} \tag{2}$$

Following [Adams et al., 2010], we let $\alpha(j) = \lambda^j \alpha_0$, for $\alpha_0 > 0$ and $\lambda \in (0, 1)$. This defines a sequence of trees $(\Pi^{(t)} = ((\pi_\epsilon^{(t)}), (\phi_{\epsilon i}^{(t)})), t \in \mathcal{T})$.

Intuitively, the prior distribution over a tree at time t is given by the posterior distribution of the (stationary) TSSBP, conditioned on the observations in some window $t-h, \ldots, t-1$. The following theorem gives the equivalence of dynamic TSSBP (dTSSBP) and TSSBP

Theorem 1. *The marginal posterior distribution of the dTSSBP, at time t, follows a TSSBP.*

Proof. The exchangeable distribution over partitions associated with a Dirichlet process is described using the Ewen's Sampling Formula (ESF). As shown by [Caron et al., 2007], the resulting random partition still follows an ESF if, at time t we deterministically delete observations from time $t-h$. The associated posterior random measure at time t will therefore follow a Dirichlet process, following de Finetti's theorem.

This result extends trivially to the dTSSBP. The child nodes in a tree are distributed according to a Dirichlet process, and maintain Dirichlet process marginals under the described deletion scheme. The posterior probabilities associated with internal nodes are distributed according to a beta distribution; the beta distribution is a special case of the Dirichlet process (where the base measure is atomic with support in two locations), therefore the marginal distribution under deletion remains a beta distribution. □

The proof is a straightforward extension of that for the generalized Pólya urn dependent Dirichlet process [Caron et al., 2007]. The above theorem implies that Equation 2 defines a *dependent tree-structured stick-breaking process.*

We note that an alternative choice for inducing dependency would be to down-weight the contribution of observations for previous time-steps. For example, we could exponentially decay the contributions of observations from previous time-steps, inducing a similar form of dependency as that found in the recurrent Chinese restaurant process [Ahmed and Xing, 2008]. However, unlike the method described in Equation 2, such an approach would not yield stationary TSSBP-distributed marginals.

3.2 Dependent hierarchical clustering

The construction above gives a distribution over infinite-dimensional trees, which in turn have a probability distribution over their nodes. In order to use this distribution in a hierarchical Bayesian model for data, we must associate each node with a parameter value $\theta_\epsilon^{(t)}$. We wish to capture two properties: 1) Within a tree $\Pi^{(t)}$, nodes have similar values to their parents;

and 2) Between trees $\Pi^{(t)}$ and $\Pi^{(t+1)}$, corresponding nodes $\epsilon^{(t)}$ and $\epsilon^{(t+1)}$ have similar values. This form of variation is shown in the bottom row of Figure 1. In this subsection, we present two models that exhibit these properties: One appropriate for real-valued data, and one appropriate for multinomial data.

3.2.1 A time-varying, tree-structured mixture of Gaussians

An infinite mixture of Gaussians is a flexible choice for density estimation and clustering real-valued observations. Here, we suggest a time-varying hierarchical clustering model that is similar to the generalized Gaussian model of [Adams et al., 2010]. The model assumes Gaussian-distributed data at each node, and allows the means of clusters to evolve in an auto-regressive model, as below:

$$\begin{aligned} \theta_{\emptyset}^{(t)} | \theta_{\emptyset}^{(t-1)} &\sim \mathcal{N}(\theta_{\emptyset}^{(t-1)}, \sigma_0 \sigma_1^a \mathbf{I}) \\ \theta_{\epsilon \cdot i}^{(t)} | \theta_{\epsilon}^{(t)}, \theta_{\epsilon \cdot i}^{(t-1)} &\sim \mathcal{N}(m, s^2 \mathbf{I}), \end{aligned} \tag{3}$$

where

$$\begin{aligned} s^2 =& \left(\frac{1}{\sigma_0 \sigma_1^{|\epsilon \cdot i|}} + \frac{1}{\sigma_0 \sigma_1^{|\epsilon \cdot i|+a}} \right)^{-1} \\ m =& s^2 \cdot \left(\frac{\theta_{\epsilon}^{(t)}}{(\sigma_0 \sigma_1^{|\epsilon \cdot i|})^2} + \frac{\eta \theta_{\epsilon \cdot i}^{(t-1)}}{\sigma_0 \sigma_1^{|\epsilon \cdot i|+a}} \right), \end{aligned}$$

$\sigma_0 > 0$, $\sigma_1 \in (0, 1)$, $\eta \in [0, 1)$, and $a \geq 1$. We denote by $\Theta^{(t)}$ the set of all parameters $\theta_{\epsilon}^{(t)}$ associated with a tree $\Pi^{(t)}$. We note that, due to self-conjugacy of the Gaussian distribution this corresponds to a Markov network with factor potentials given by unnormalized Gaussian distributions: Up to a normalizing constant, the factor potential associated with the link between $\theta_{\epsilon}^{(t-1)}$ and $\theta_{\epsilon}^{(t)}$ is Gaussian with variance $\sigma_0 \sigma_1^{|\epsilon|}$, and the factor potential associated with the link between $\theta_{\epsilon}^{(t)}$ and $\theta_{\epsilon \cdot i}^{(t)}$ is Gaussian with variance $\sigma_0 \sigma_1^{|\epsilon \cdot i|+a}$

3.2.2 A time-varying model for hierarchically clustering documents

Given a dictionary of V words, a document can be represented using a V-dimensional term frequency (TF) or term frequency-inverse document frequency (TF-IDF) vector, that corresponds to a location on the surface of the $(V-1)$-dimensional unit sphere. The von Mises-Fisher distribution, with mean direction $\boldsymbol{\mu}$ and concentration parameter τ, provides a distribution on this space. A mixture of von Mises-Fisher distributions can, therefore, be used to cluster documents [Banerjee et al., 2005, Gopal and Yang, 2014]. Following the terminology of topic modeling [Blei et al., 2003], the mean direction $\boldsymbol{\mu}_k$ associated with the kth cluster can be interpreted as the topic associated with that cluster.

We construct a time-dependent hierarchical clustering model appropriate for documents by associating each node of our dependent nonparametric tree with such a topic. Concretely, let $\mathbf{x}_n^{(t)}$ be the vector associated with the nth document at time t. We assign a mean parameter $\theta_{\epsilon}^{(t)}$ to each node ϵ in each tree $\Pi^{(t)}$ as

$$\begin{aligned}\theta_{\emptyset}^{(t)}|\theta_{\emptyset}^{(t-1)} &\sim \text{vMF}(\tau_{\emptyset}^{(t)}, \rho_{\emptyset}^{(t)}) \\ \theta_{\epsilon\cdot i}^{(t)}|\theta_{\epsilon}^{(t)}, \theta_{\epsilon\cdot i}^{(t-1)} &\sim \text{vMF}(\tau_{\epsilon\cdot i}^{(t)}, \rho_{\epsilon\cdot i}^{(t)}),\end{aligned} \tag{4}$$

where

$$\begin{aligned}\rho_{\emptyset}^{(t)} =& \kappa_0\sqrt{1+\kappa_1^{2a}+2\kappa_1^a(\theta_{-1}^{(t)}\cdot\theta_{\emptyset}^{(t-1)})} \\ \tau_{\emptyset}^{(t)} =& \frac{\kappa_0\theta_{-1}^{(t)}+\kappa_0\kappa_1^a\theta_{\emptyset}^{(t-1)}}{\rho_{\emptyset}^{(t)}} \\ \rho_{\epsilon\cdot i}^{(t)} =& \kappa_0\kappa_1^{|\epsilon\cdot i|}\sqrt{1+\kappa_1^{2a}+2\kappa_1^a(\theta_{\epsilon}^{(t)}\cdot\theta_{\epsilon\cdot i}^{(t-1)})} \\ \tau_{\epsilon\cdot i}^{(t)} =& \frac{\kappa_0\kappa_1^{|\epsilon\cdot i|}\theta_{\epsilon}^{(t)}+\kappa_0\kappa_1^{|\epsilon\cdot i|+a}\theta_{\epsilon\cdot i}^{(t-1)}}{\rho_{\epsilon\cdot i}^{(t)}},\end{aligned}$$

$\kappa_0 > 0$, $\kappa_1 > 1$, and $\theta_{-1}^{(t)}$ is a probability vector of the same dimension as the $\theta_{\epsilon}^{(t)}$ that can be interpreted as the parent of the root node at time t.[1] This yields similar dependency behavior to that described in Section 3.2.1.

Conditioned on $\Pi^{(t)}$ and $\Theta^{(t)} = (\theta_{\epsilon}^{(t)})$, we sample each document $\mathbf{x}_n^{(t)}$ according to

$$\begin{aligned}z_n^{(t)} \sim& \text{Discrete}(\Pi^{(t)}) \\ \mathbf{x}_n \sim& \text{vMF}(\theta^{(t)}, \beta)\end{aligned} \tag{5}$$

This is a hierarchical extension of the temporal vMF mixture proposed by [Gopal and Yang, 2014].

4 Online Learning

In many time-evolving applications, we observe data points in an online setting. We are typically interested in obtaining predictions for future data points, or characterizing the clustering structure of current data, rather than improving predictive performance on historic data. We therefore propose a sequential online learning algorithm, where at each time t we infer the parameter settings for the tree $\Pi^{(t)}$ conditioned on the previous trees, which we do not re-learn. This allows us to focus our computational efforts on the most recent (and likely relevant) data. This has the added advantage of reducing the computational demands of the algorithm, as we do not incorporate a backwards pass through the data, and are only ever considering a fraction of the data at a time.

In developing an inference scheme, there is always a trade-off between estimate quality and computational requirements. MCMC samplers are often the "gold standard" of inference techniques, because they have the true posterior distribution as the stationary distribution of their Markov Chain. However, they can be very slow, particularly in complex models. Estimating the parameter setting that maximizes the data likelihood is a much cheaper, but cannot capture the full posterior.

[1] In our experiments, we set $\theta_{-1}^{(t)}$ to be the average over all data points at time t. This ensures that the root node is close to the cetroid of the data, rather than the periphery.

In order to develop an inference algorithm that is parallelizable, runs in reasonable time, but still obtains good predictive performance, we combine Gibbs sampling steps for learning the tree parameters ($\Pi^{(t)}$) and the topic indicators ($z_n^{(t)}$) with a MAP method for estimating the location parameters ($\theta_\epsilon^{(t)}$). The resulting algorithm has the following desirable properties:

1. The priors for $\nu_\epsilon^{(t)}, \psi_\epsilon^{(t)}$ only depend on $\{z_n^{(0)}\} \dots \{z_n^{(t-1)}\}$, whose sufficient statistics $\{X_\epsilon^{(0)}, Y_\epsilon^{(0)}\}$ $\dots \{X_\epsilon^{(t-1)}, Y_\epsilon^{(t-1)}\}$ can be updated in amortized constant time.
2. The posteriors for $\nu_\epsilon^{(t)}, \psi_\epsilon^{(t)}$ are conditionally independent given $\{z_n^{(1)}\} \dots \{z_n^{(t)}\}$. Hence we can Gibbs sample $\nu_\epsilon^{(t)}, \psi_\epsilon^{(t)}$ in parallel given the cluster assignments $\{z_n^{(1)}\} \dots \{z_n^{(t)}\}$ (or more precisely, their sufficient statistics $\{X_\epsilon, Y_\epsilon\}$). Similarly, we can Gibbs sample the cluster/topic assignments $\{z_n^{(t)}\}$ in parallel given the parameters $\{\nu_\epsilon^{(t)}, \psi_\epsilon^{(t)}, \theta_\epsilon^{(t)}\}$ and the data, as well as infer the MAP estimate of $\{\theta_\epsilon^{(t)}\}$ in parallel given the data and the cluster/topic assignments. Because of the online assumption, we do not consider evidence from times $u > t$.

Sampling $\nu_\epsilon^{(t)}, \psi_\epsilon^{(t)}$ Due to the conjugacy between the beta and binomial distributions, we can easily Gibbs sample the stick-breaking parameters

$$\nu_\epsilon^{(t)} | X_\epsilon, Y_\epsilon \sim \text{Beta}\big(1 + \textstyle\sum_{t'=t-h}^{t} X_\epsilon^{(t')}, \alpha(|\epsilon|) + \sum_{t'=t-h}^{t} Y_\epsilon^{(t')}\big)$$
$$\psi_{\epsilon\cdot i}^{(t)} | X_{\epsilon\cdot i}, Y_{\epsilon\cdot i} \sim \text{Beta}\big(1 + \textstyle\sum_{t'=t-h}^{t} (X_{\epsilon\cdot i}^{(t')} + Y_{\epsilon\cdot i}^{(t')}), \gamma + \sum_{j>i} \sum_{t'=t-h}^{t} (X_{\epsilon\cdot j}^{(t')} + Y_{\epsilon\cdot j}^{(t')})\big).$$

The $\nu_\epsilon^{(t)}, \psi_\epsilon^{(t)}$ distributions for each node are conditionally independent given the counts X, Y, and so the sampler can be parallelized. We only explicitly store $\pi_\epsilon^{(t)}, \phi_\epsilon^{(t)}, \theta_\epsilon^{(t)}$ for nodes ϵ with nonzero counts, i.e. $\sum_{t'=t-h}^{t} X_\epsilon^{(t')} + Y_\epsilon^{(t')} > 0$.

Sampling $z_n^{(t)}$ Conditioned on the $\nu_\epsilon^{(t)}$ and $\psi_\epsilon^{(t)}$, the distribution over the cluster assignments $z_n^{(t)}$ is just given by the TSSBP. We therefore use the slice sampling method described in [Adams et al., 2010] to Gibbs sample $z_n^{(t)} \mid \{\nu_\epsilon^{(t)}\}, \{\psi_\epsilon^{(t)}\}, x_n^{(t)}, \theta$. Since the cluster assignments are conditionally independent given the tree, this step can be performed in parallel.

Learning θ It is possible to Gibbs sample the cluster parameters θ; however, in the document clustering case described in Section 3.2.2, this requires far more time than sampling all other parameters. To improve the speed of our algorithm, we instead use *maximum a posteriori* (MAP) estimates for θ, obtained using a parallel coordinate ascent algorithm. Notably, conditioned on the trees at time $t-1$ and $t+1$, the $\theta_\epsilon^{(t)}$ for odd-numbered tree depths $|\epsilon|$ are conditionally independent given the $\theta_{\epsilon'}^{(t)}$s at even-numbered tree depths $|\epsilon'|$, and vice versa. Hence, our algorithm alternates between parallel optimization of odd-depth $\theta_\epsilon^{(t)}$, and parallel optimization of even-depth $\theta_\epsilon^{(t)}$.

In general, the conditional distribution of a cluster parameter $\theta_\epsilon^{(t)}$ depends on the values of its predecessor $\theta_\epsilon^{(t-1)}$, its postdecessor $\theta_\epsilon^{(t+1)}$, its parent at time t, and its children at time t. In some cases, not all of these values will be available – for example if a node was unoccupied at previous time steps. In this case, the distribution now depends on the full history of the parent node. For computational reasons, and because we do not wish to store the full history,

we approximate the distribution as being dependent only on observed members of the node's Markov blanket.

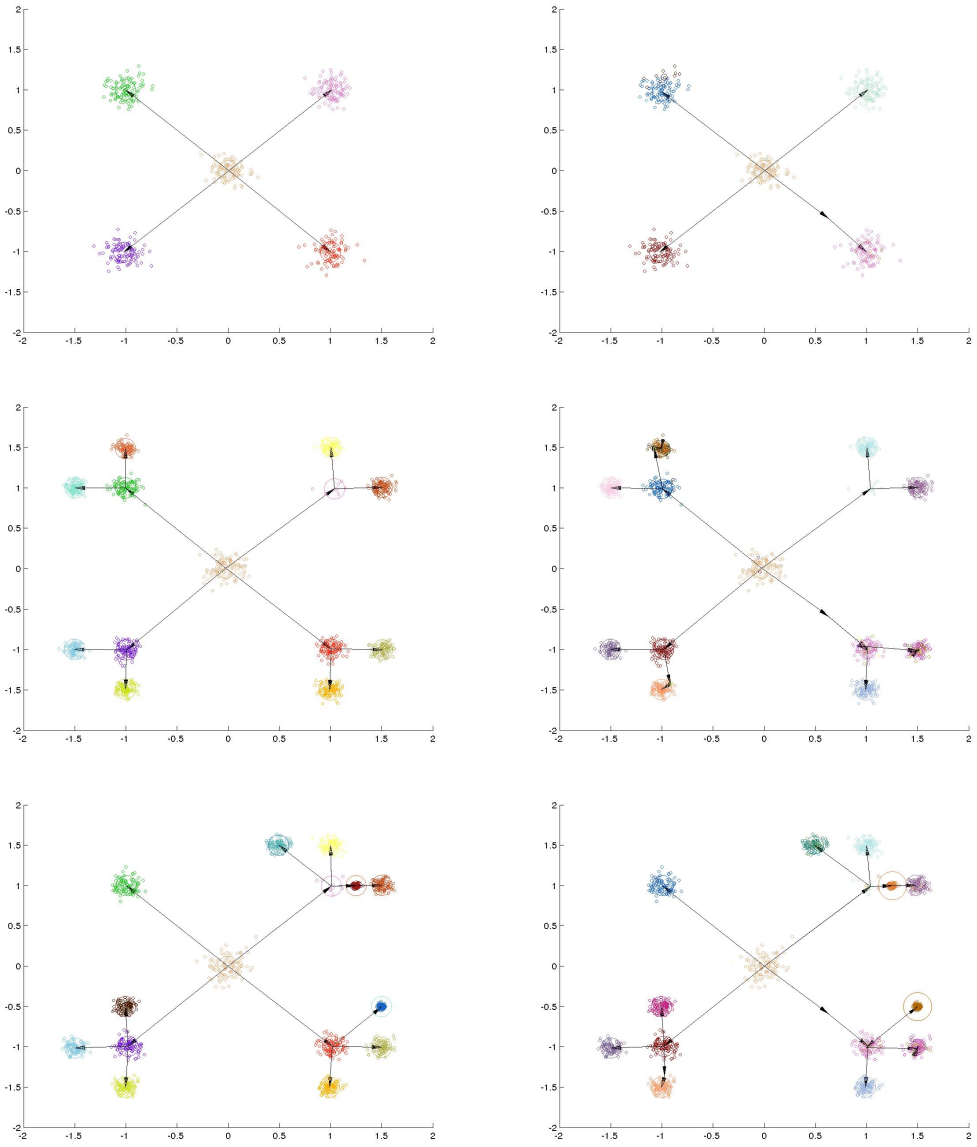

Figure 2: **Left: Ground truth tree**, evolving over three time steps and **Right: Recovered tree structure**, over three consecutive time periods. Each row represents an epoch. Each color indicates a node in the tree and each arrow indicates a branch connecting parent to child; nodes are consistently colored across time.

5 Experimental evaluation

We evaluate the performance of our model on both synthetic and real-world data sets. Evaluation on synthetic data sets allows us to verify that our inference algorithm allows us to recover the "true" evolving hierarchical structure underlying our data. Evaluation on real-world data allows us to evaluate whether our modeling assumptions are useful in practice.

5.1 Synthetic data

We manually created a time-evolving tree, as shown in Figure 2 left, with Gaussian-distributed data at each node. This synthetic time-evolving tree features temporal variation in node probabilities, temporal variation in node parameters, and addition and deletion of nodes. Using the Gaussian model described in Equation 3, we inferred the structure of the tree at each time period as described in Section 4. Figure 2 right shows the recovered tree structure, demonstrating the ability of our inference algorithm to recover the expected evolving hierarchical structure. Note that it accurately captures evolution in node probabilities and location, and the addition and deletion of new nodes.

5.2 Real-world data

In Section 3.2.2, we described how the dependent TSSBP can be combined with a von Mises-Fisher likelihood to cluster documents. To evaluate this model, we looked at three corpora:

- TWITTER: 673,102 tweets containing hashtags relevant to the NFL, collected over 18 weeks in 2011 and containing 2,636 unique words (after stopwording). We grouped the tweets into 9 two-week epochs.
- PNAS: 79,800 paper titles from the Proceedings of the National Academy of Sciences between 1915 and 2005, containing 36,901 unique words (after stopwording). We grouped the titles into 10 ten-year epochs.
- STATE OF THE UNION (SOU): Presidential SoU addresses from 1790 through 2002, containing 56,352 sentences and 21,505 unique words (after stopwording). We grouped the sentences into 21 ten-year epochs.

In each case, documents were represented using their vector of term frequencies.

Our hypothesis is that the topical structure of language is *hierarchically structured* and *time-evolving*, and that a model that captures these properties will achieve better performance than models that ignore hierarchical structure and/or temporal evolution. To test these hypotheses, we compare our dependent tree-structured stick-breaking process (dTSSBP) against several online nonparametric models for document clustering:

1. Multiple tree-structured stick-breaking process (T-TSSBP): We modeled the entire corpus using the stationary TSSBP model, with each node modeled using an independent von Mises-Fisher distribution. Each time period is modeled with a separate tree, using a similar implementation to our time-dependent TSSBP.
2. "Online" tree-structured stick-breaking processes (o-TSSBP): This simulates online learning of a single, stationary tree over the entire corpus. We used our dTSSBP implementation with an infinite window $h = \infty$, and once a node is created at time t, we prevent its vMF mean $\theta_\epsilon^{(t)}$ from changing in future time points.

3. Dependent Dirichlet process (dDP): We modeled the entire corpus using an h-order Markov generalized Pólya urn DDP [Caron et al., 2007]. This model was implemented by modifying our dTSSBP code to have a single level. Node parameters were evolved as $\theta_k^{(t)} \sim \text{vMF}(\theta_k^{(t)}, \xi)$.
4. Multiple Dirichlet process (T-DP): We modeled the entire corpus using DP mixtures of von Mises-Fisher distributions, one DP per time period. Each node was modeled using an independent von Mises-Fisher distribution. We used our own implementation.
5. "Online" Dirichlet process (o-DP): This simulates online learning of a single DP over the entire corpus. We used our dDP implementation with an infinite window $h = \infty$, and once a cluster is instantiated at time t, we prevent its vMF mean $\theta^{(t)}$ from changing in future time points.

Evaluation scheme: We divide each dataset into two parts: the first 50%, and last 50% of time points. We use the first 50% to tune model parameters and select a good random restart (by training on 90% and testing on 10% of the data at each time point), and then use the last 50% to evaluate the performance of the best parameters/restart (again, by training on 90% and testing on 10% data). When training the 3 TSSBP-based models, we grid-searched $\kappa_0 \in \{1, 10, 100, 1000, 10000\}$, and fixed $\kappa_1 = 1$, $a = 0$ for simplicity. Each value of κ_0 was run 5 times to get different random restarts, and we took the best κ_0-restart pair for evaluation on the last 50% of time points. For the 3 DP-based models, there is no κ_0 parameter, so we simply took 5 random restarts and used the best one for evaluation. For all TSSBP- and DP-based models, we repeated the evaluation phase 5 times to get error bars. Every dTSSBP trial completed in < 20 minutes on a single processor core, while we observed moderate (though not perfectly linear) speedups with 2-4 processors.

Parameter settings: For all models, we estimated each node/cluster's vMF concentration parameter β from the data. For the TSSBP-based models, we used stick breaking parameters $\gamma = 0.5$ and $\alpha(d) = 0.5^d$, and set $\theta_{-1}^{(t)}$ to the average document term frequency vector at time t. In order to keep running times reasonable, we limit the TSSBP-based models to a maximum depth of either 3 or 4 (we report results for both)[2]. For the DP-based models, we used a Dirichlet process concentration parameter of 1. The dDP's inter-epoch vMF concentration parameter was set to $\xi = 0.001$.

Results: Table 1 shows the average log (unnormalized) likelihoods on the test sets (from the last 50% of time points). The tree-based models uniformly out-perform the non-hierarchical models, while the max-depth-4 tree models outperform the max-depth-3 ones. On all 3 datasets, the max-depth-4 dTSSBP uniformly outperforms all models.

Discussion We started of with the assumption that evolving hierarchical clustering is a better way to model data generated over a period of time. Since all hierarchical models ($\{d, o, T\}$-TSSBP) have better test log-likelihood as compared to non-hierarchical models ($\{d, o, T\}$-DP) we conclude that hierarchies of clusters are important while modeling text data. Within hierarchical models, we see that the model (d-TSSBP) that capture evolving hierarchies performs significantly better than stationary models; thus, making a compelling case to use evolving hierarchies to model time varying data.

[2]One justification is that shallow hierarchies are easier to interpret than deep ones; see [Blei et al., 2004, Ho et al., 2012].

	dTSSBP		o-TSSBP		T-TSSBP	
Depth limit	4	3	4	3	4	3
Twitter	$\mathbf{522 \pm 4.35}$	249 ± 0.98	414 ± 3.31	199 ± 2.19	335 ± 54.8	182 ± 24.1
SoU	$\mathbf{2708 \pm 32.0}$	1320 ± 33.6	1455 ± 44.5	583 ± 16.4	1687 ± 329	1089 ± 143
PNAS	$\mathbf{4562 \pm 116}$	3217 ± 195	2672 ± 357	1163 ± 196	4333 ± 647	2962 ± 685

	dDP	o-DP	T-DP
Twitter	204 ± 8.82	136 ± 0.42	112 ± 10.9
SoU	834 ± 51.2	633 ± 18.8	890 ± 70.5
PNAS	2374 ± 51.7	1061 ± 10.5	2174 ± 134

Table 1: Test set average log-likelihood on three datasets.

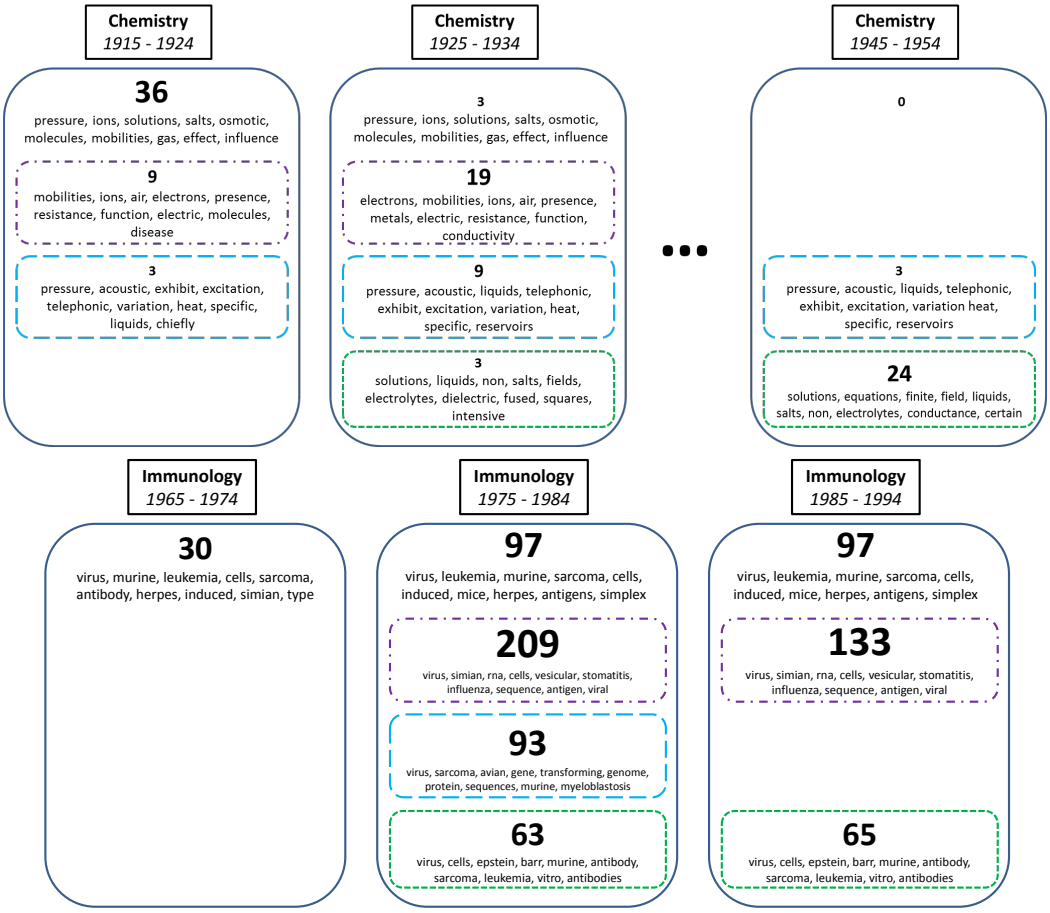

Figure 3: **PNAS dataset:** Birth, growth, and death of tree-structured topics in our dTSSBP model. This illustration captures some trends in American scientific research throughout the 20th century, by focusing on the evolution of parent and child topics in two major scientific areas: Chemistry and Immunology (the rest of the tree has been omitted for clarity). At each epoch, we show the number of documents assigned to each topic, as well as it's most popular words (according to the vMF mean θ).

5.3 Qualitative results with Discussions

In addition to better quantitative results, we find that the time-dependent tree model gives good qualitative performance. Figure 3 shows two time-evolving sub-trees obtained from the

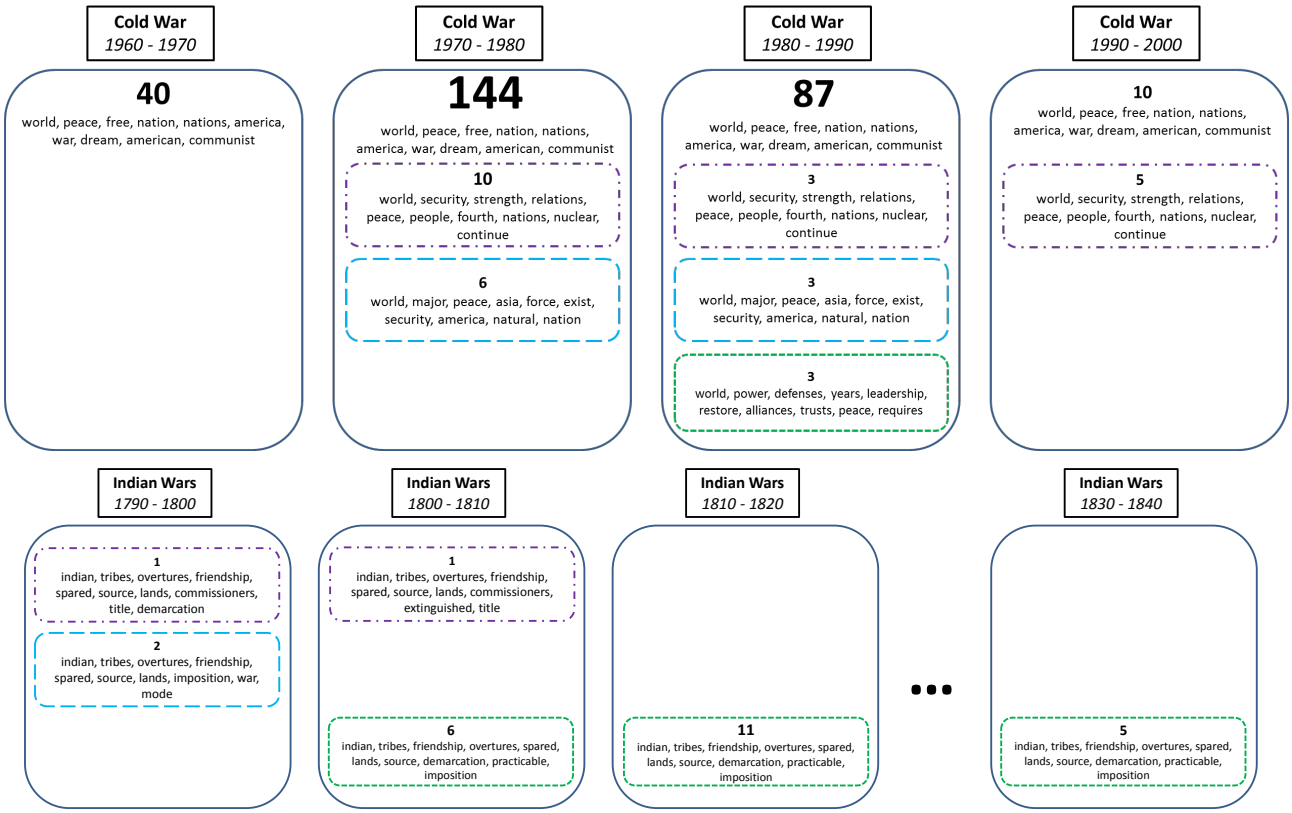

Figure 4: **State of the Union dataset:** Birth, growth, and death of tree-structured topics in our dTSSBP model. This illustration captures some key events in American history. At each epoch, we show the number of documents assigned to each topic, as well as it's most popular words (according to the vMF mean θ).

PNAS data set. The top level shows a sub-tree concerned with Chemistry; the bottom level shows a sub-tree concerned with Immunology. Our dynamic tree model discovers closely-related topics and groups them under a sub-tree, and creates, grows and destroys individual sub-topics as needed to fit the data. For instance, our model captures the sudden surge in Immunology-related research from 1975-1984, which happened right after the structure of the antibody molecule was identified a few years prior.

In the Chemistry topic, the study of mechanical properties of materials (pressure, acoustic properties, specific heat, etc) has a continued presence throughout the century. The study of electrical properties of materials starts off with a topic (in purple) that seems to be devoted to Physical Chemistry. However, following the development of Quantum Mechanics in the 30s, this line of research became more closely aligned with Physics than Chemistry, and it disappears from the sub-tree. In its wake, we see the growth of a topic more concerned with electrolytes, solutions and salts, which remained the within the sphere of Chemistry.

Figure 4 shows time-evolving sub-trees obtained from the State of the Union dataset. We see a sub-tree tracking the development of the Cold War. The parent node contains general terms relevant to the Cold War; starting from the 1970s, a child node (shown in purple)

contains terms relevant to nuclear arms control, in light of the Strategic Arms Limitation Talks of that decade. The same decade also sees the birth of a child node focused on Asia (shown in cyan), contemporaneous with President Richard Nixon's historic visit to China in 1972. In addition to the Cold War, we also see topics corresponding to events such as the Mexican War, the Civil War and the Indian Wars, demonstrating our model's ability to detect events in a timeline.

6 Future Work

One of the drawbacks of having parameters pass down a tree in the manner discussed in section 3.2.2 is that children nodes may have a lot of similarity among themselves; thereby, leading to multiple similar topics taking birth at the same time. We believe this can be solved by using a penalty that ensures diversity among children and leave it for future work. Another interesting direction would be to explore large scale implementation of our model which gives near linear speedup. Also, our current model assumes that each item belongs to a single cluster. In future, it would be interesting to explore ways to handle admixtures where we relax the assumption so that an item can belong to multiple clusters. In the future it will also be interesting to see whether we can extend the parallel inference framework of [Williamson et al., 2013] and [Dubey et al., 2014b] to do exact parallel inference in our proposed model.

Acknowledgements This work was done with Qirong Ho, Sinead Williamson and Eric P. Xing and was accepted at NIPS 2014 [Dubey et al., 2014a]. A approximate distributed extension to the paper was also published in ICML 2015 [Hu et al., 2015].

References

[Adams et al., 2010] Adams, R., Ghahramani, Z., and Jordan, M. (2010). Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems*.

[Ahmed and Xing, 2008] Ahmed, A. and Xing, E. (2008). Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In *SDM*.

[Banerjee et al., 2005] Banerjee, A., Dhillon, I., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382.

[Blei and Frazier, 2011] Blei, D. and Frazier, P. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12(2461–2488).

[Blei et al., 2004] Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*.

[Blei and Lafferty, 2006] Blei, D. and Lafferty, J. (2006). Dynamic topic models. In *International Conference on Machine Learning*.

[Blei et al., 2003] Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[Caron et al., 2007] Caron, F., Davy, M., and Doucet, A. (2007). Generalized Polya urn for time-varying Dirichlet processes. In *uai*.

[Dubey et al., 2013] Dubey, A., Hefny, A., Williamson, S., and Xing, E. (2013). A non-parametric mixture model for topic modeling over time. In *SDM*.

[Dubey et al., 2014a] Dubey, A., Ho, Q., Williamson, S., and Xing, E. (2014a). Dependent nonparametric trees for dynamic hierarchical clustering. In *NIPS*.

[Dubey et al., 2014b] Dubey, A., Williamson, S., and Xing, E. (2014b). Parallel markov chain monte carlo for pitman-yor mixture models. In *UAI 2014*.

[Ferguson, 1973] Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.

[Gopal and Yang, 2014] Gopal, S. and Yang, Y. (2014). Von Mises-Fisher clustering models. In *International Conference on Machine Learning*.

[Ho et al., 2012] Ho, Q., Eisenstein, J., and Xing, E. (2012). Document hierarchies from text and links. In *Proceedings of the 21st international conference on World Wide Web*, pages 739–748. ACM.

[Hu et al., 2015] Hu, Z., Ho, Q., Dubey, A., and Xing, E. (2015). Large-scale distributed dependent nonparametric trees. In *ICML*.

[Kingman, 1982] Kingman, J. (1982). On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43.

[Lin et al., 2010] Lin, D., Grimson, E., and Fisher, J. (2010). Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in Neural Information Processing Systems*.

[MacEachern, 1999] MacEachern, S. N. (1999). Dependent nonparametric processes. In *Bayesian Statistical Science*.

[Neal, 2003] Neal, R. M. (2003). Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629.

[Pitman and Yor, 1997] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900.

[Rodriguez et al., 2008] Rodriguez, A., Dunson, D., and Gelfand, A. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483).

[Sudderth and Jordan, 2008] Sudderth, E. and Jordan, M. (2008). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems*.

[Wang and McCallum, 2006] Wang, X. and McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. In *Knowledge Discovery and Data Mining*.

[Williamson et al., 2013] Williamson, S., Dubey, A., and Xing, E. (2013). Parallel markov chain monte carlo for nonparametric mixture models. In *ICML*.

[Williamson et al., 2010] Williamson, S., Orbanz, P., and Ghahramani, Z. (2010). Dependent Indian buffet processes. In *Artificial Intelligence and Statistics*.