A Nested HDP for Hierarchical Topic Models

John Paisley
Department of EECS
UC Berkeley

Chong Wang
Dept. of Machine Learning
Carnegie Mellon University

David Blei
Dept. of Computer Science
Princeton University

Michael I. Jordan
Department of EECS
UC Berkeley

Abstract

We develop a nested hierarchical Dirichlet process (nHDP) for hierarchical topic modeling. The nHDP is a generalization of the nested Chinese restaurant process (nCRP) that allows each word to follow its own path to a topic node according to a document-specific distribution on a shared tree. This alleviates the rigid, single-path formulation of the nCRP, allowing a document to more easily express thematic borrowings as a random effect. We demonstrate our algorithm on 1.8 million documents from *The New York Times*.[1]

1 Introduction

Organizing things hierarchically is a natural process of human activity. A hierarchical tree-structured representation of data can provide an illuminating means for understanding and reasoning about the information it contains. The nested Chinese restaurant process (nCRP) [2] is a model that performs this task for the problem of topic modeling. It does this by learning a tree structure for the underlying topics, with the inferential goal being that topics closer to the root are more general, and gradually become more specific in thematic content when following a path down the tree.

The nCRP is limited in the hierarchies it can model. We illustrate this limitation in Figure 1. The nCRP models the topics that go into constructing a document as lying along one path of the tree. This significantly limits the underlying structure that can be learned from the data. The nCRP performs *document-specific* path clustering; our goal is to develop a related Bayesian nonparametric prior that performs *word-specific* path clustering. We illustrate this objective in Figure 1. To this end, we make use of the hierarchical Dirichlet process [3], developing a novel prior that we refer to as the *nested hierarchical Dirichlet process* (nHDP). With the nHDP, a top-level nCRP becomes a base distribution for a collection of second-level nCRPs, one for each document. The nested HDP provides the opportunity for cross-thematic borrowing that is not possible with the nCRP.

2 Nested Hierarchical Dirichlet Processes

Nested CRPs. Nested Chinese restaurant processes (nCRP) are a tree-structured extension of the CRP that are useful for hierarchical topic modeling. A natural interpretation of the nCRP is as a tree where each parent has an infinite number of children. Starting from the root node, a path is traversed down the tree. Given the current node, a child node is selected with probability proportional to the previous number of times it was selected among its siblings, or a new child is selected with probability proportional to a parameter $\alpha > 0$.

For hierarchical topic modeling with the nCRP, each node has an associated discrete distribution on word indices drawn from a Dirichlet distribution. Each document selects a path down the tree according to a Markov process, which produces a sequence of topics. A stick-breaking process provides a distribution on the selected topics, after which words for a document are generated by first drawing a topic, and then drawing the word index.

HDPs. The HDP is a multi-level version of the Dirichlet process. It makes use of the idea that the base distribution for the DP can be discrete. A discrete base distribution allows for multiple draws from the DP prior to place probability mass on the same subset of atoms. Hence different groups of data can share the same atoms, but place different probability distributions on them. The HDP draws its discrete base from a DP prior, and so the atoms are learned.

[1] This is a workshop version of a longer paper. Please see [1] for more details, including a scalable inference algorithm.

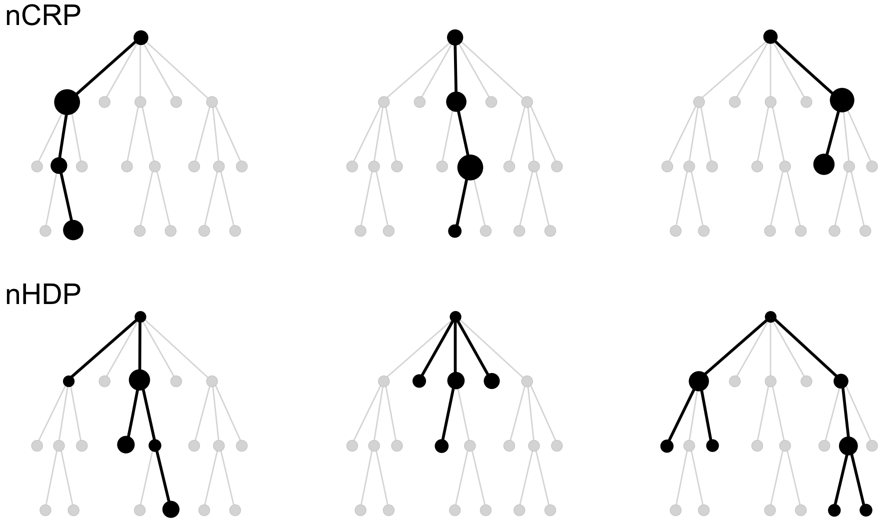

Figure 1: An example of path structures for the nested Chinese restaurant process (nCRP) and the nested hierarchical Dirichlet process (nHDP) for hierarchical topic modeling. With the nCRP, the topics for a document are restricted to lying along a single path to a root node. With the nHDP, each document has access to the entire tree, but a document-specific distribution on paths will place high probability on a particular subtree. The goal of the nHDP is to learn a thematically consistent tree as achieved by the nCRP, while allowing for the cross-thematic borrowings that naturally occur within a document.

Nested HDPs. In building on the nCRP framework, our goal is to allow for each document to have access to the entire tree, while still learning document-specific distributions on topics that are thematically coherent. Ideally, each document will still exhibit a dominant path corresponding to its main themes, but with offshoots allowing for random effects. Our two major changes to the nCRP formulation toward this end are that (i) each word follows its own path to a topic, and (ii) each document has its own distribution on paths in a shared tree.

With the nHDP, all documents share a global nCRP. This nCRP is equivalently an infinite collection of Dirichlet processes with a transition rule between DPs starting from a root node. With the nested HDP, we use each Dirichlet process in the global nCRP as a base for a second-level DP drawn independently for each document. This constitutes an HDP. Using a nesting of HDPs allows each document to have its own tree where the transition probabilities are defined over the same subset of nodes, but where the values of these probabilities are document-specific.

For each node in the tree, we also generate document-specific switching probabilities that are i.i.d. beta random variables. When walking down the tree for a particular word, we stop with the switching probability of the current node, or continue down the tree according to the probabilities of the child HDP. We summarize this process in Algorithm 1.

Sample Results. We present some qualitative results for the nHDP topic model on a set of 1.8 million documents from *The New York Times*. These results were obtained using a scalable variational inference algorithm [4] after one pass through the data set. In Figure 2 we show example topics from the model and their relative structure. We show four topics from the top level of the tree (shaded), and connect topics according to parent/child relationship. The model learns a meaningful hierarchical structure; for example, the sports subtree branches into the various sports, which themselves appear to branch by teams. In the foreign affairs subtree, children tend to group by major subregion and then branch out into subregion or issue.

Algorithm 1 Generating Documents with the Nested Hierarchical Dirichlet Process

```
Step 1. Generate a global tree by constructing an nCRP.
Step 2. For each document, generate a document tree from the HDP and switching probabilities for each node.
Step 3. Generate a document. For word $n$ in document $d$,
        a) Walking down the tree using HDPs. At current node, continue/stop according to its switching probability.
        b) Sample word $n$ from the discrete distribution at the terminal node.
```

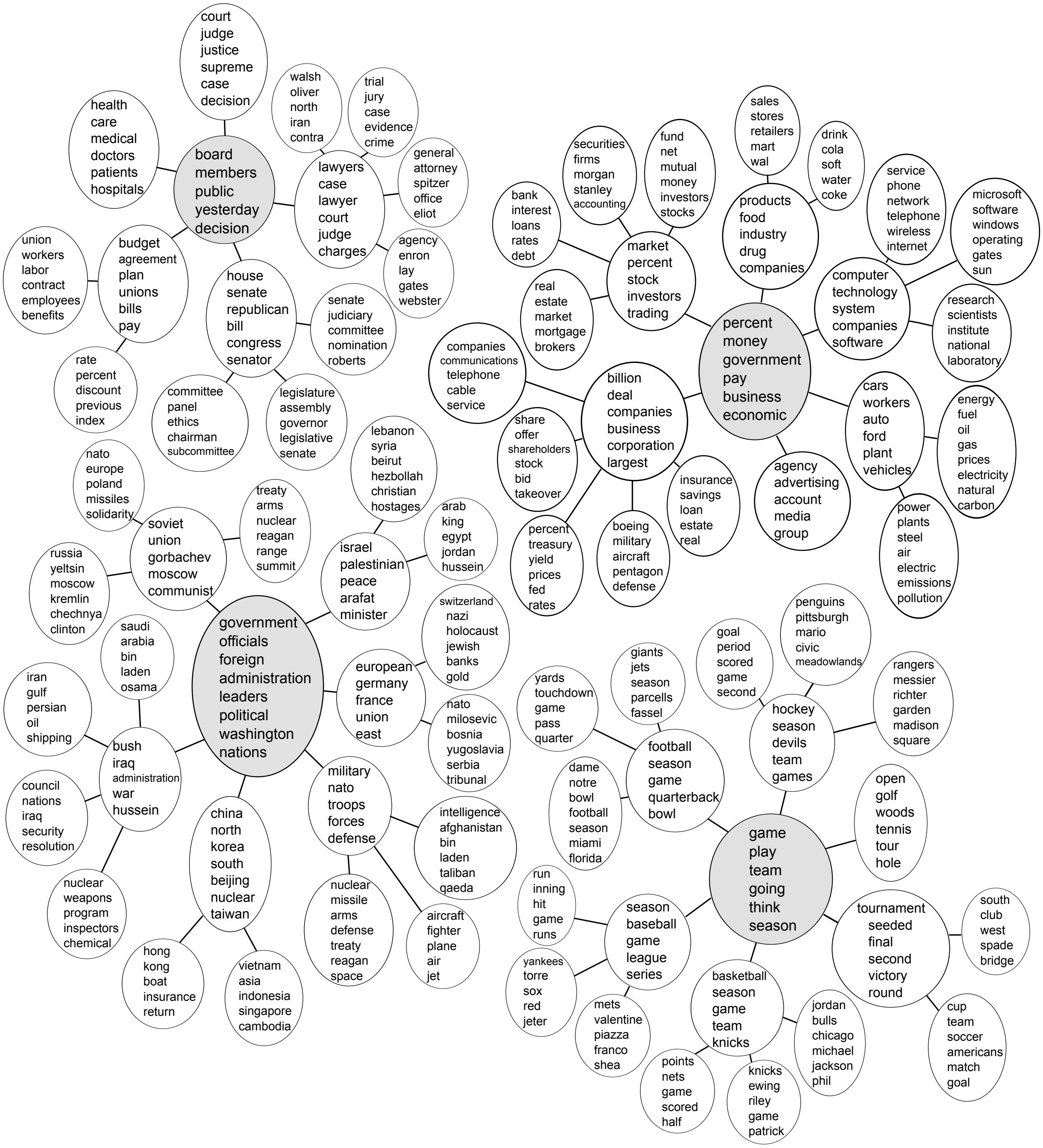
court judge justice supreme case decision
health care medical doctors patients hospitals
board members public yesterday decision
walsh oliver north iran contra
trial jury case evidence crime
lawyers case lawyer court judge charges
general attorney spitzer office eliot
agency enron lay gates webster
union workers labor contract employees benefits
budget agreement plan unions bills pay
house senate republican bill congress senator
senate judiciary committee nomination roberts
rate percent discount previous index
committee panel ethics chairman subcommittee
legislature assembly governor legislative senate
nato europe poland missiles solidarity
soviet union gorbachev moscow communist
treaty arms nuclear reagan range summit
lebanon syria beirut hezbollah christian hostages
arab king egypt jordan hussein
israel palestinian peace arafat minister
russia yeltsin moscow kremlin chechnya clinton
government officials foreign administration leaders political washington nations
switzerland nazi holocaust jewish banks gold
saudi arabia bin laden osama
european germany france union east
nato milosevic bosnia yugoslavia serbia tribunal
iran gulf persian oil shipping
bush iraq administration war hussein
council nations iraq security resolution
china north korea south beijing nuclear taiwan
military nato troops forces defense
intelligence afghanistan bin laden taliban qaeda
nuclear weapons program inspectors chemical
nuclear missile arms defense treaty reagan space
aircraft fighter plane air jet
hong kong boat insurance return
vietnam asia indonesia singapore cambodia
sales stores retailers mart wal
securities firms morgan stanley accounting
fund net mutual money investors stocks
drink cola soft water coke
service phone network telephone wireless internet
microsoft software windows operating gates sun
bank interest loans rates debt
products food industry drug companies
market percent stock investors trading
real estate market mortgage brokers
computer technology system companies software
research scientists institute national laboratory
percent money government pay business economic
companies communications telephone cable service
billion deal companies business corporation largest
cars workers auto ford plant vehicles
energy fuel oil gas prices electricity natural carbon
share offer shareholders stock bid takeover
insurance savings loan estate real
agency advertising account media group
power plants steel air electric emissions pollution
percent treasury yield prices fed rates
boeing military aircraft pentagon defense
penguins pittsburgh mario civic meadowlands
goal period scored game second
giants jets season parcells fassel
yards touchdown game pass quarter
rangers messier richter garden madison square
hockey season devils team games
football season game quarterback bowl
dame notre bowl football season miami florida
open golf woods tennis tour hole
game play team going think season
run inning hit game runs
season baseball game league series
yankees torre sox red jeter
tournament seeded final second victory round
south club west spade bridge
mets valentine piazza franco shea
basketball season game team knicks
jordan bulls chicago michael jackson phil
cup team soccer americans match goal
points nets game scored half
knicks ewing riley game patrick

Figure 2: Tree-structured topics from The New York Times. A shaded node is a top-level node.

References

[1] J. Paisley, C. Wang, D. Blei, and M. I. Jordan, “Nested hierarchical Dirichlet processes,” *arXiv:1210.6738*, 2012.

[2] D. Blei, T. Griffiths, and M. Jordan, “The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies,” *Journal of the ACM*, vol. 57, no. 2, pp. 7:1–30, 2010.

[3] Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical Dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[4] M. Hoffman, D. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” *arXiv:1206.7051*, 2012.