The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan
JACM, January 2010

Michael Bryant
Jixiong Wang

Motivation

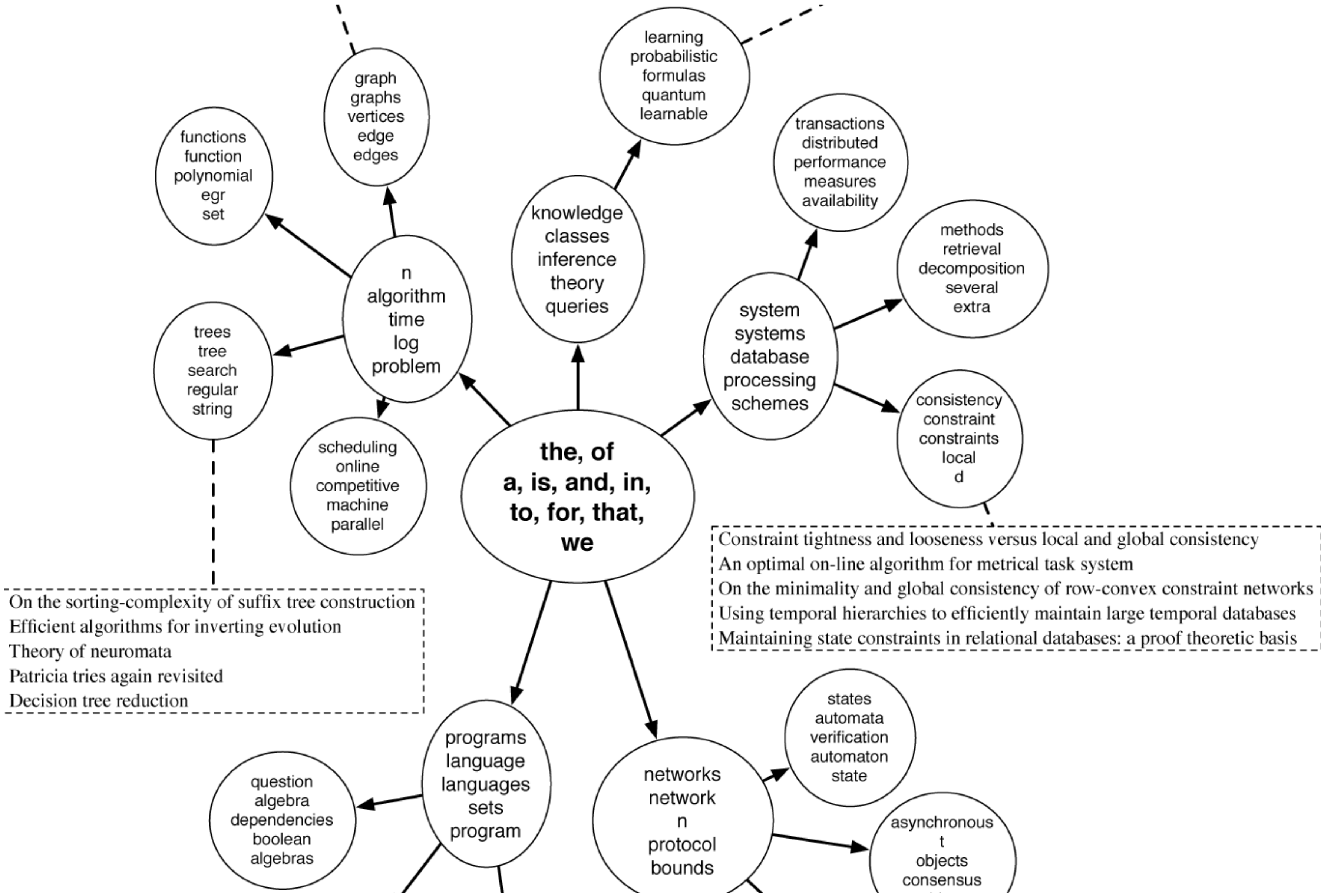
Problem:

- To learn topic models for collections of text, images and other semi-structured corpora.

- The original topic models treat topics as a “flat” set of probability distributions, with no direct relationship between one topic and another. They fail to indicate the level of abstraction of a topic, or how the various topics are related.

- We want an algorithm to both find useful sets of topics and learn to organize the topics according to a hierarchy in which more abstract topics are near the root of the hierarchy and more concrete topics are near the leaves.

Slide3



the, of
a, is, and, in,
to, for, that,
we
n
algorithm
time
log
problem
functions
function
polynomial
egr
set
graph
graphs
vertices
edge
edges
trees
tree
search
regular
string
scheduling
online
competitive
machine
parallel
knowledge
classes
inference
theory
queries
learning
probabilistic
formulas
quantum
learnable
system
systems
database
processing
schemes
transactions
distributed
performance
measures
availability
methods
retrieval
decomposition
several
extra
consistency
constraint
constraints
local
d
Constraint tightness and looseness versus local and global consistency
An optimal on-line algorithm for metrical task system
On the minimality and global consistency of row-convex constraint networks
Using temporal hierarchies to efficiently maintain large temporal databases
Maintaining state constraints in relational databases: a proof theoretic basis
On the sorting-complexity of suffix tree construction
Efficient algorithms for inverting evolution
Theory of neuromata
Patricia tries again revisited
Decision tree reduction
programs
language
languages
sets
program
question
algebra
dependencies
boolean
algebras
networks
network
n
protocol
bounds
states
automata
verification
automaton
state
asynchronous
t
objects
consensus

Motivation

Problem:

- To learn topic models for collections of text, images and other semi-structured corpora.

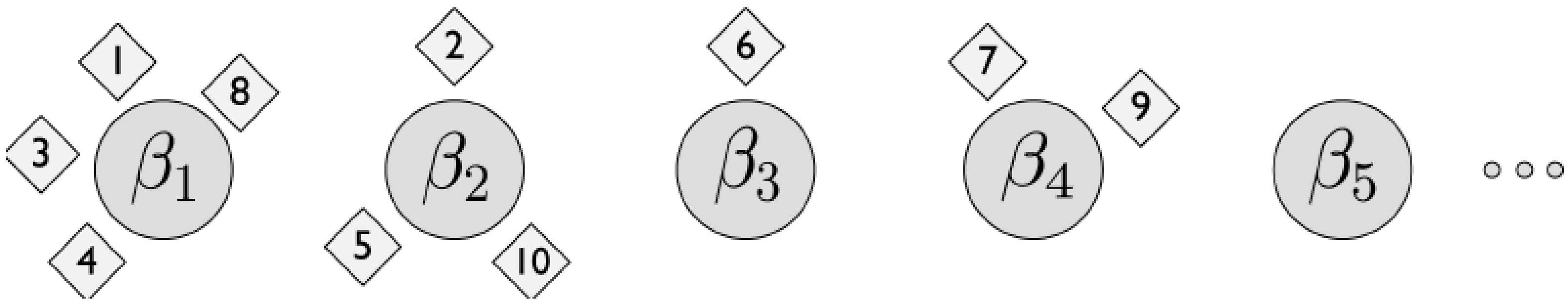
- The original topic models treat topics as a “flat” set of probability distributions, with no direct relationship between one topic and another. They fail to indicate the level of abstraction of a topic, or how the various topics are related.

- We want an algorithm to both find useful sets of topics and learn to organize the topics according to a hierarchy in which more abstract topics are near the root of the hierarchy and more concrete topics are near the leaves.

- While a classical unsupervised analysis might require the topology of the hierarchy to be chosen in advance, we want the approach to place high probability on those hierarchies that best explain the data. We need a distribution on topologies.

- We wish to allow this distribution to have its support on arbitrary topologies, -- there should be no limitations such as a maximum depth or maximum branching factor. BNP is needed.

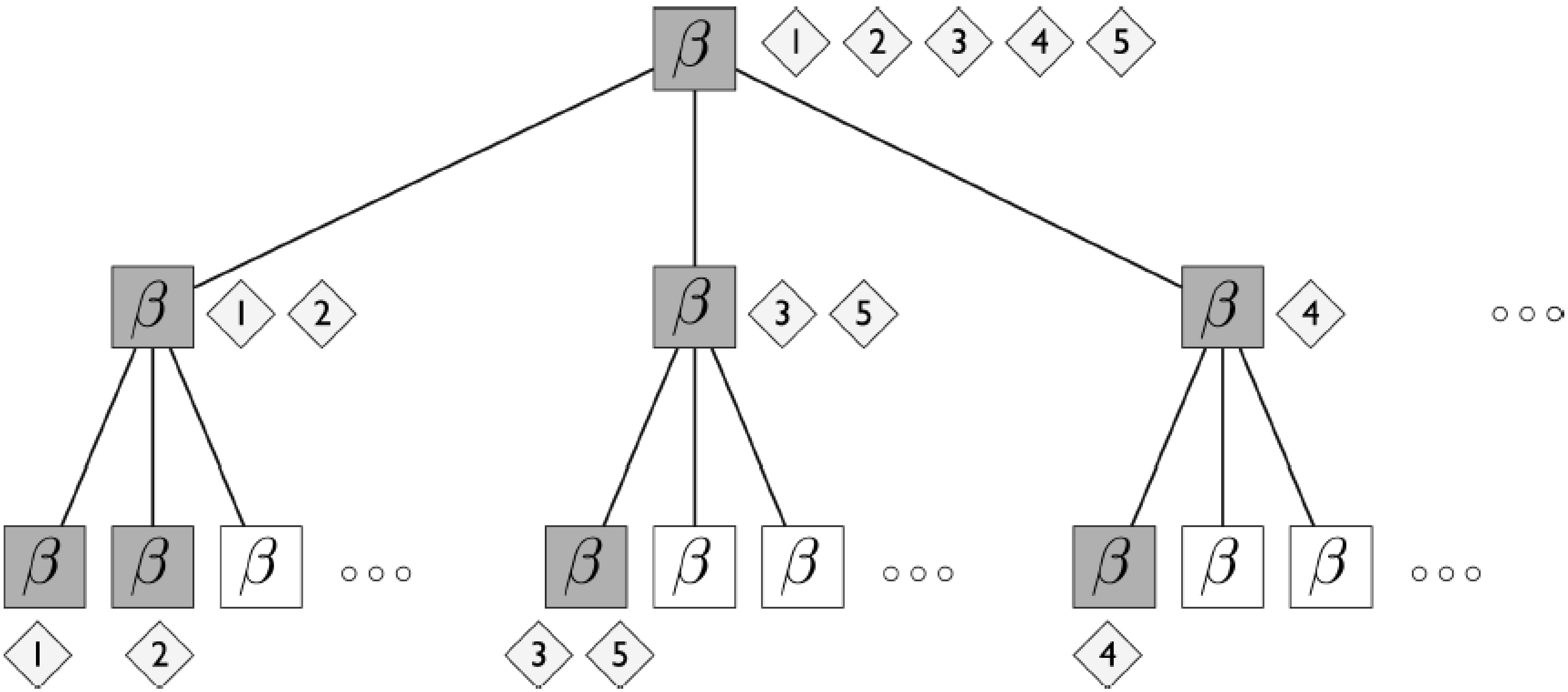
Chinese Restaurant Process



$$p(\text{occupied table } i \mid \text{ previous customers}) = \frac{n_i}{\gamma+n-1}$$

$$p(\text{next unoccupied table} \mid \text{ previous customers}) = \frac{\gamma}{\gamma+n-1}$$

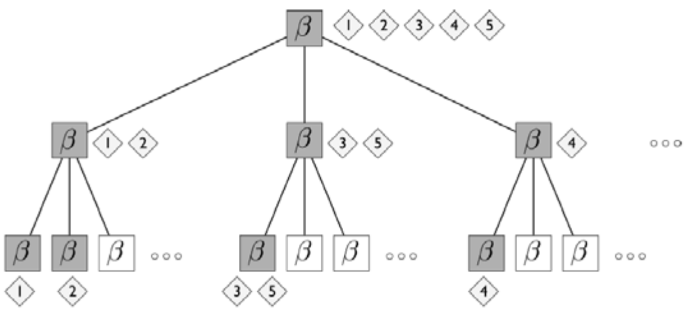
Nested Chinese Restaurant Process

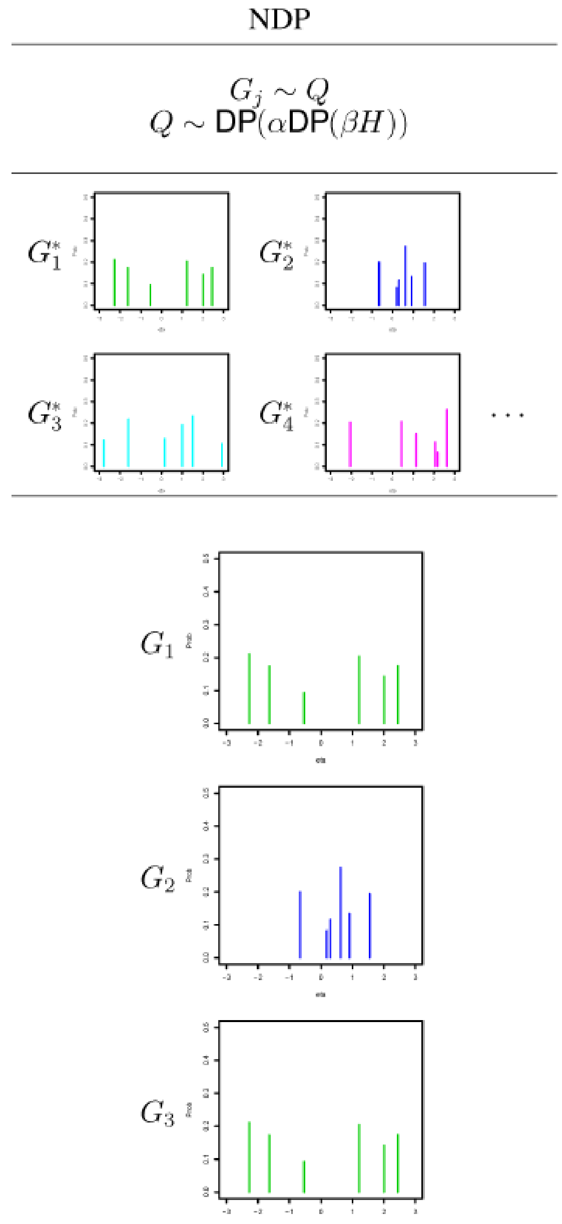


$$p(\text{occupied table } i \mid \text{previous customers}) = \frac{n_i}{\gamma+n-1}$$

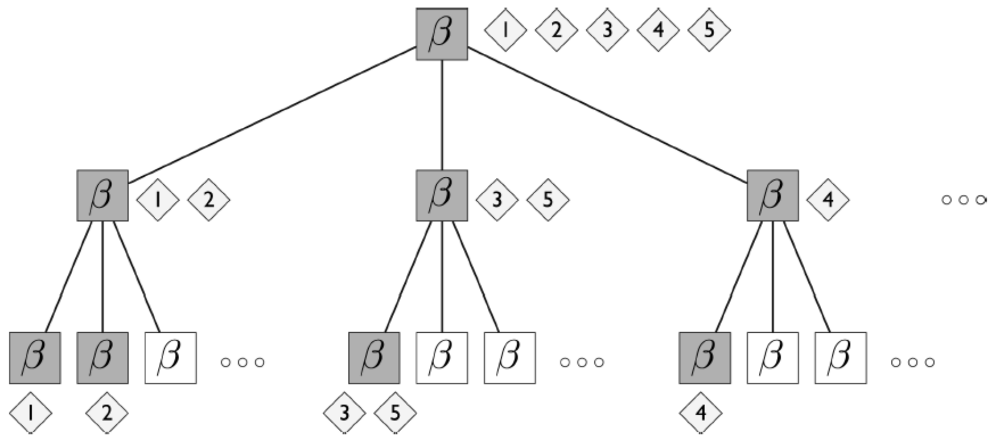
$$p(\text{next unoccupied table} \mid \text{previous customers}) = \frac{\gamma}{\gamma+n-1}$$

nCRP vs nDP


β
1 2 3 4 5
1 2
3 5
4
3 5
4


NDP
$G_j \sim Q$
$Q \sim \mathrm{DP}(\alpha \mathrm{DP}(\beta H))$
G_1^*
G_2^*
G_3^*
G_4^*
G_1
G_2
G_3

Hierarchical Latent Dirichlet Allocation



for each table $k \in \mathcal{T}$ in the infinite tree **do**
- Draw a topic $\beta_k \sim \text{Dirichlet}(\eta)$

end for

for each document $d \in \{1, 2, \ldots, D\}$ **do**
- Draw $\boldsymbol{c}_d \sim \text{nCRP}(\gamma)$
- Draw a distribution over levels in the tree, $\theta_d | \{m, \pi\} \sim \text{GEM}(m, \pi)$

    for each word in document d **do**
  - Choose level $Z_{d,n} | \theta_d \sim \text{Discrete}(\theta_d)$
  - Choose word $W_{d,n} | \{z_{d,n}, \boldsymbol{c}_d, \boldsymbol{\beta}\} \sim \text{Discrete}(\beta_{c_d}[z_{d,n}])$

    end for

end for

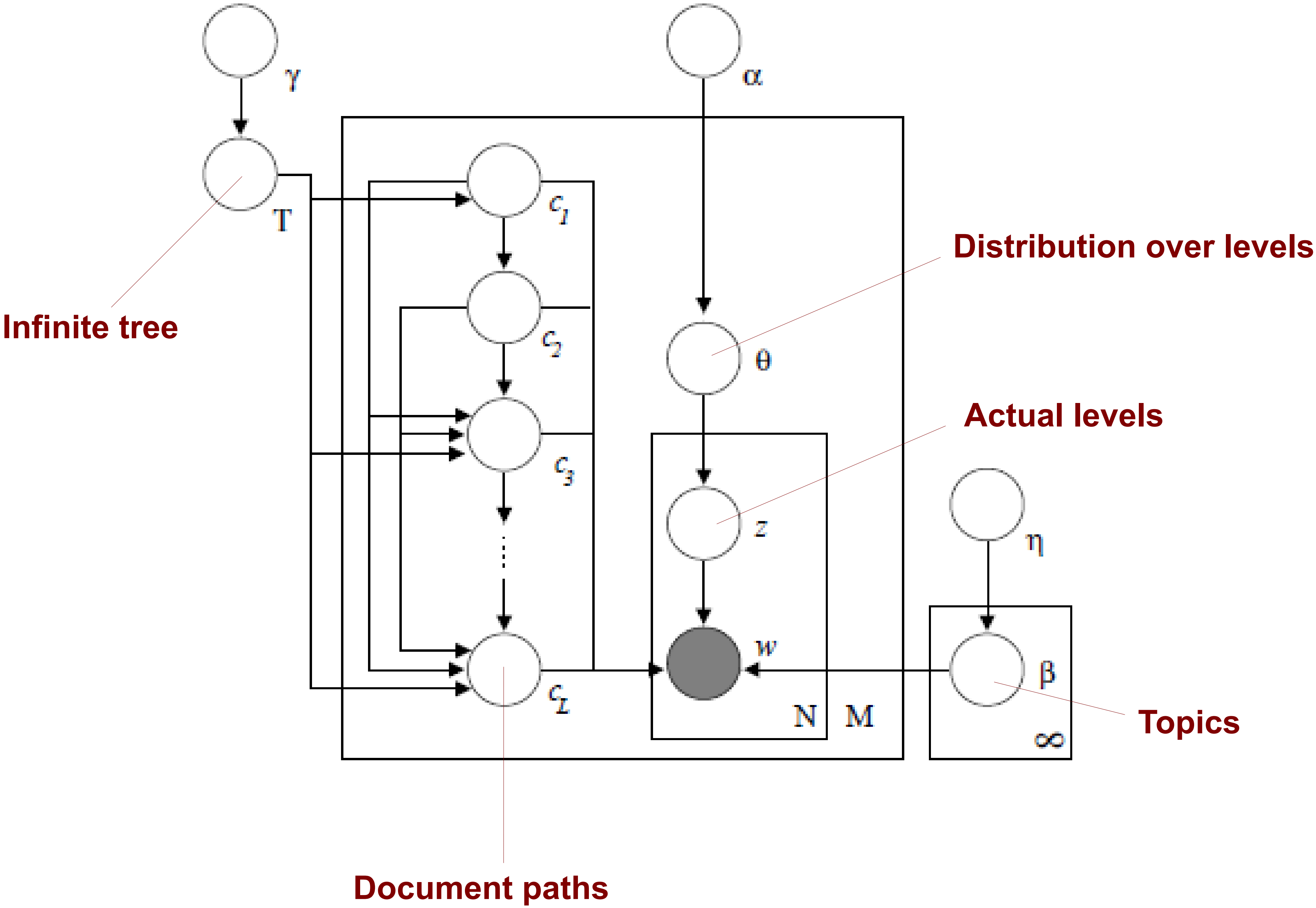
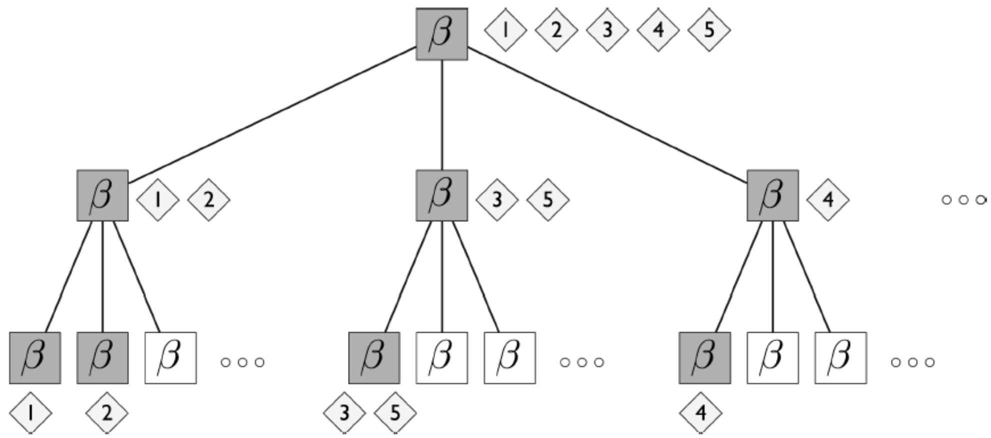


Figure from Blei, et al 2003

Hierarchical Latent Dirichlet Allocation



for each table $k \in \mathcal{T}$ in the infinite tree **do**
  - Draw a topic $\beta_k \sim \text{Dirichlet}(\eta)$

end for

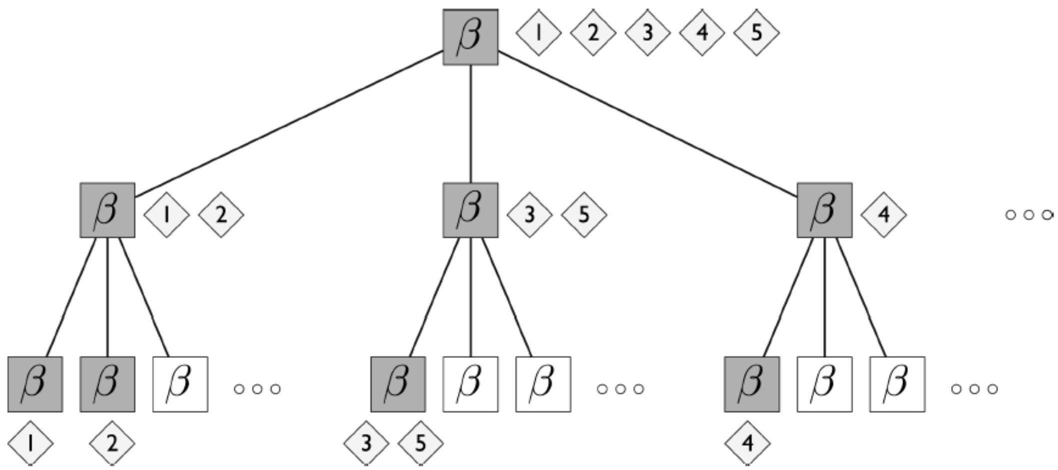
for each document $d \in \{1, 2, \ldots, D\}$ **do**
  - Draw $\boxed{c_d} \sim \text{nCRP}(\gamma)$
  - Draw a distribution over levels in the tree, $\theta_d|\{m, \pi\} \sim \text{GEM}(m, \pi)$

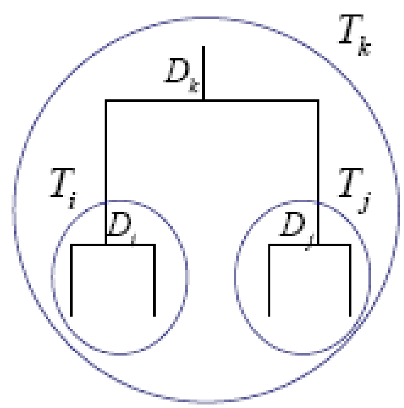
  for each word in document d **do**
  - Choose level $\boxed{Z_{d,n}}|\theta_d \sim \text{Discrete}(\underline{\theta_d})$
  - Choose word $W_{d,n}|\{z_{d,n}, c_d, \beta\} \sim \text{Discrete}(\underline{\beta_{c_d}}[z_{d,n}])$

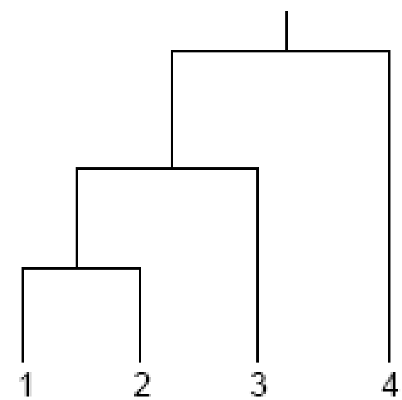
  end for

end for

hLDA vs Hierarchical Clustering

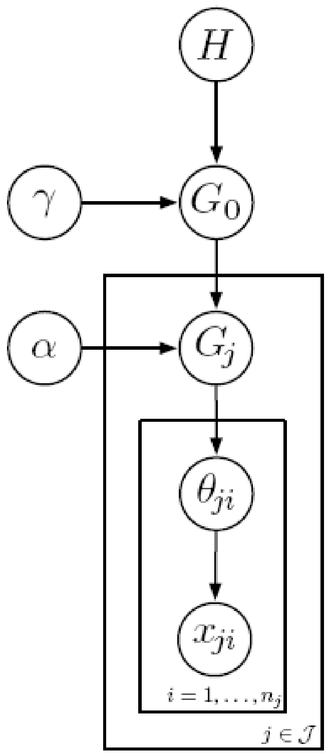


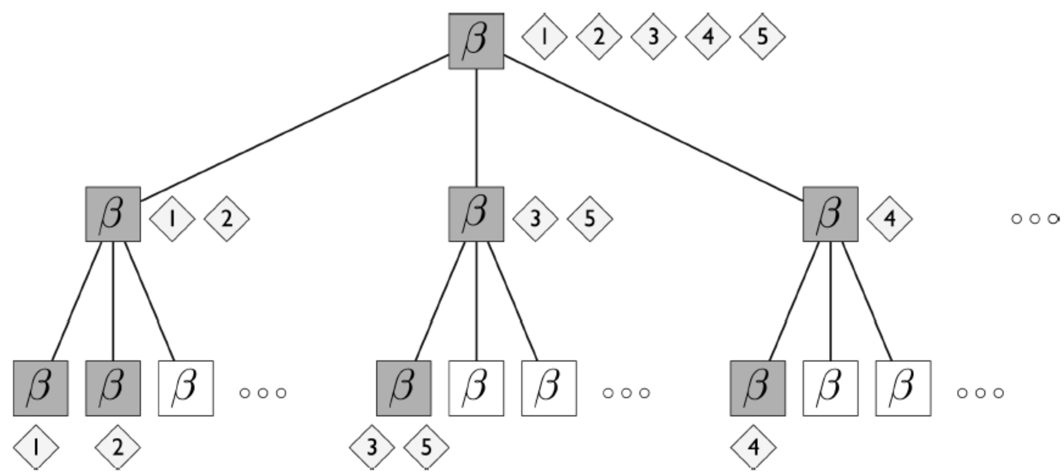




hLDA vs HDP-LDA

$$
\begin{aligned}
G_0 \mid \gamma, H &\sim \mathrm{DP}(\gamma, H) \\
G_j \mid \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) && \text{for each document } j \in \mathcal{J} \\
\theta_{ji} \mid G_j &\sim G_j && \text{for each word } i = 1, \ldots, n_j \\
x_{ji} \mid \theta_{ji} &\sim F_{\theta_{ji}},
\end{aligned}
$$





Notation

$\mathbf{c}_{1:D}$ = Paths for documents 1,...,D

$\mathbf{z}_{1:D}$ = Level assignments for documents 1,...,D

$\mathbf{z}_{-(d,n)}$ = All level assignments excluding that for word *n* in doc *d*

$\mathbf{z}_{d,-n}$ = All level assignments for doc *d* excluding word *n*

γ = nCRP hyperparameter

η = Topic hyperparameter

m, π = Level distribution hyperparameters

$\mathbf{w}_{1:D}$ = Words for documents 1,...,D

Inference for hLDA

- Given the model and a corpus, we want to estimate the distribution of (some of) the latent variables. In math words:

$$p(\mathbf{c}_{1:D}, \mathbf{z}_{1:D} | \gamma, \eta, m, \pi, \mathbf{w}_{1:D})$$

- Approximate this posterior by *Collapsed Gibbs Sampling*.

- Iterate between:

(1) Sampling level assignments:

$$p(z_{d,n} | \mathbf{z}_{-(d,n)}, \mathbf{c}, \mathbf{w}, m, \pi, \eta) \propto p(z_{d,n} | \mathbf{z}_{d,-n}, m, \pi) p(w_{d,n} | \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta)$$

(2) Sampling paths:

$$p(\mathbf{c}_d | \mathbf{w}, \mathbf{c}_{-d}, \mathbf{z}, \eta, \gamma) \propto p(\mathbf{c}_d | \mathbf{c}_{-d}, \gamma) p(\mathbf{w}_d | \mathbf{c}, \mathbf{w}_{-d}, \mathbf{z}, \eta)$$

Inference (CGS)

Integrate out distributions over levels and topics

Sample the paths and level assignments

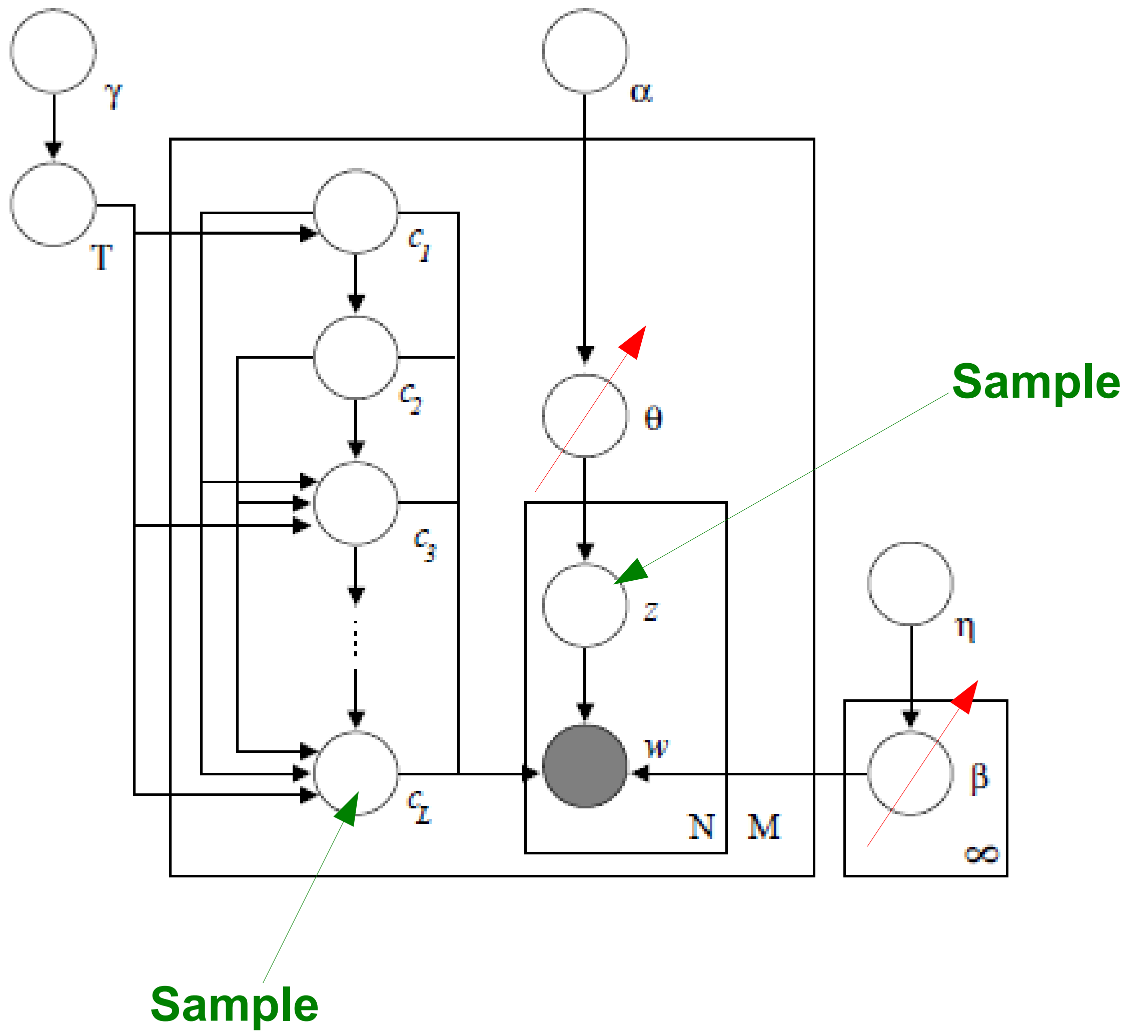
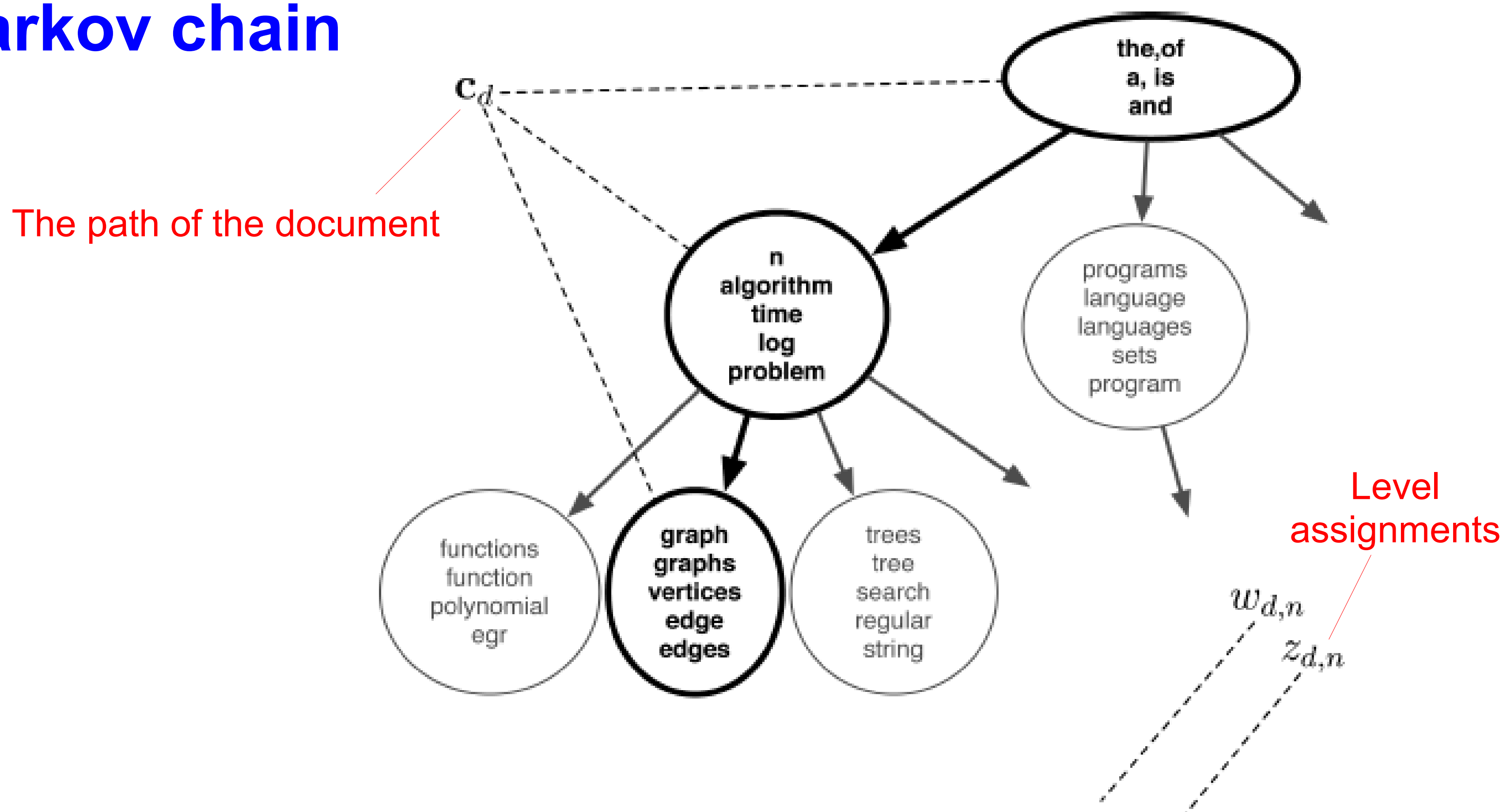


Figure from Blei, et al 2003

The Markov chain



All$_0$ previously$_1$ known$_0$ efficient$_2$ maximum-flow algorithms$_1$ work$_1$ by$_0$ finding$_1$ augmenting paths$_1$, either$_1$ one$_0$ path$_0$ at$_0$ a$_0$ time$_1$ (as$_0$ in$_0$ the$_0$ original$_2$ Ford and$_0$ Fulkerson algorithm$_1$) or$_0$ all$_0$ shortest-length augmenting paths$_1$ at$_0$ once$_0$ (using$_0$ the$_0$ layered network$_2$ approach$_1$ of$_0$ Dinic). An$_0$ alternative$_1$ method$_0$ based$_0$ on$_0$ the$_0$ preflow concept$_0$ of$_0$ Karzanov is$_0$ introduced$_0$. A$_0$ preflow is$_0$ like$_0$ a$_0$ flow$_2$ except$_1$ that$_0$ the$_0$ total$_0$ amount$_0$ flowing into$_1$ a$_0$ vertex$_2$ is$_0$ allowed$_0$ to$_0$ exceed the$_0$ total$_0$ amount$_0$ flowing out$_2$. The$_0$ method$_0$ maintains a$_0$ preflow in$_0$ the$_0$ original$_2$ network$_2$ and$_0$ pushes local$_0$ flow$_2$ excess toward$_1$ the$_0$ sink along$_0$ what$_0$ are$_0$ estimated to$_0$ be$_0$ shortest$_2$ paths$_1$. The$_0$ algorithm$_1$ and$_0$ its$_0$ analysis$_0$ are$_0$ simple$_1$ and$_0$ intuitive$_1$, yet$_0$ the$_0$ algorithm$_1$ runs$_0$ as$_0$ fast$_1$ as$_0$ any$_0$ other$_0$ known$_0$ method$_0$ on$_0$ dense graphs$_2$ achieving an$_0$ O(n) time$_1$ bound$_1$ on$_0$ an$_0$ n-vertex$_2$ graph$_2$ by$_0$ incorporating the$_0$ dynamic$_1$ tree$_1$ data$_0$ structure$_1$ of$_0$ Sleator and$_0$ Tarjan...

Sampling level assignments

$$p(z_{d,n}|\mathbf{z}_{-(d,n)}, \mathbf{c}, \mathbf{w}, m, \pi, \eta) \propto p(z_{d,n}|\mathbf{z}_{d,-n}, m, \pi)p(w_{d,n}|\mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta)$$

Probability of level assignment given other level assignments

Probability of word given level assignments

Infinite

Want to sample:

$$p(z_{d,n} = k|\mathbf{z}_{-(d,n)}, \mathbf{c}, \mathbf{w}, m, \pi, \eta), \qquad k = 1, 2, \ldots$$

Solution: sample in stages

- Sample a level from all currently represented levels, plus one level deeper
- If deeper level is chosen, sample from a Bernoulli to go even deeper, iteratively

Sampling level assignments

For already represented levels:

Expected length of the k-th stick

$$p(z_{d,n} = k|\mathbf{z}_{d,-n}, m, \pi) = \mathbb{E}\left[V_k \prod_{j=1}^{k-1}(1 - V_j)\right]$$

$$= \mathbb{E}[V_k] \prod_{j=1}^{k-1} \mathbb{E}[1 - V_j]$$

$$= \frac{m\pi + \#[\mathbf{z}_{d,-n} = k]}{\pi + \#[\mathbf{z}_{d,-n} \geq k]} \prod_{j=1}^{k-1} \frac{(1-m)\pi + \#[\mathbf{z}_{d,-n} > j]}{\pi + \#[\mathbf{z}_{d,-n} \geq j]}$$

$$p(w_{d,n}|\mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta) \propto \#[\mathbf{z}_{-(d,n)} = z_{d,n},\ \mathbf{c}_{z_{d,n}} = c_{d,z_{d,n}},\ \mathbf{w}_{-(d,n)} = w_{d,n}] + \eta$$

Smoothed number of similar words assigned to this level in this path

For the deeper levels:

$$p(z_{d,n} > \max(\mathbf{z}_{d,-n})|\mathbf{z}_{d,-n}, \mathbf{w}, m, \pi, \eta) = 1 - \sum_{j=1}^{\max(\mathbf{z}_{d,-n})} p(z_{d,n} = j|\mathbf{z}_{d,-n}, \mathbf{w}, m, \pi, \eta)$$

Leftover probability after we account for the already represented levels

Sampling level assignments

If we choose to go deeper...

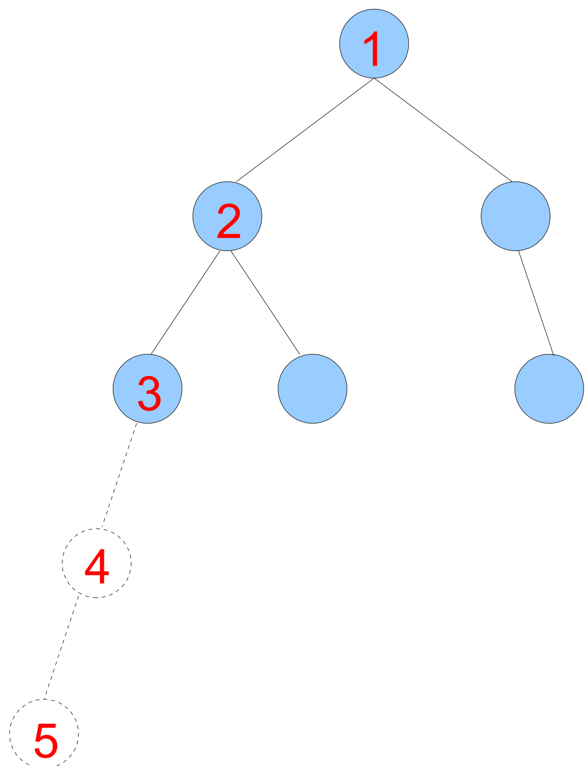
Start with:

$$\ell = \max(\mathbf{z}_{d,-n}) + 1$$

And sample from a Bernoulli with parameters:

$$\begin{aligned} p(z_{d,n} = \ell \,|\, z_{d,-n}, z_{d,n} > \ell - 1, \mathbf{w}, m, \pi, \eta) &= (1-m)p(w_{d,n} \,|\, \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta) \\ p(z_{d,n} > \ell \,|\, z_{d,-n}, z_{d,n} > \ell - 1) &= 1 - p(z_{d,n} = \ell \,|\, z_{d,-n}, z_{d,n} > \ell - 1, \mathbf{w}, m, \pi, \eta). \end{aligned}$$

Until we get a “hit”.



Sampling paths

We're only concerned with paths of the length of the deepest level assignment variable for the document under consideration.

$$p(\mathbf{c}_d|\mathbf{w},\mathbf{c}_{-d},\mathbf{z},\eta,\gamma) \propto p(\mathbf{c}_d|\mathbf{c}_{-d},\gamma)p(\mathbf{w}_d|\mathbf{c},\mathbf{w}_{-d},\mathbf{z},\eta)$$

Probability of path given all other paths

Probability of words given path

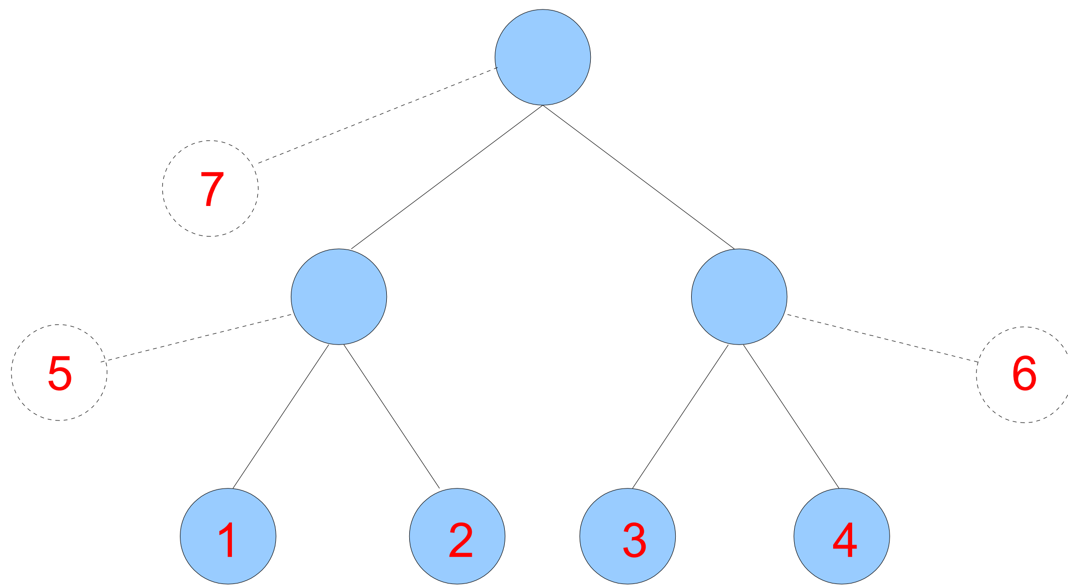
$p(\mathbf{c}_d|\mathbf{c}_{-d},\gamma)$ = nCRP prior

$$p(\mathbf{w}_d|\mathbf{c},\mathbf{w}_{-d},\mathbf{z},\eta) = \prod_{\ell=1}^{\max(\mathbf{z}_d)} \frac{\Gamma(\sum_w \#[\mathbf{z}_d=\ell,\mathbf{c}_{-d,\ell}=c_{d,\ell},\mathbf{w}_{-d}=w]+V\eta)}{\prod_w \Gamma(\#[\mathbf{z}_{-d}=\ell,\mathbf{c}_{-d,\ell}=c_{d,\ell},\mathbf{w}_{-d}=w]+\eta)} \times \frac{\prod_w \Gamma(\#[\mathbf{z}=\ell,\mathbf{c}_\ell=c_{d,\ell},\mathbf{w}=w]+\eta)}{\Gamma(\sum_w \#[\mathbf{z}=\ell,\mathbf{c}_\ell=c_{d,\ell},\mathbf{w}=w]+V\eta)}$$

Note: path must be drawn as a block; levels are not independent.

Sampling paths

All the possible paths a new document can take through the existing tree:



Sampling the hyperparameters

Take the nice Bayesian approach of putting priors on the hyperparameters:

$$\begin{aligned} m &\sim \mathrm{Beta}(\alpha_1, \alpha_2) \\ \pi &\sim \mathrm{Exponential}(\alpha_3) \\ \gamma &\sim \mathrm{Gamma}(\alpha_4, \alpha_5) \\ \eta &\sim \mathrm{Exponential}(\alpha_6) \end{aligned}$$

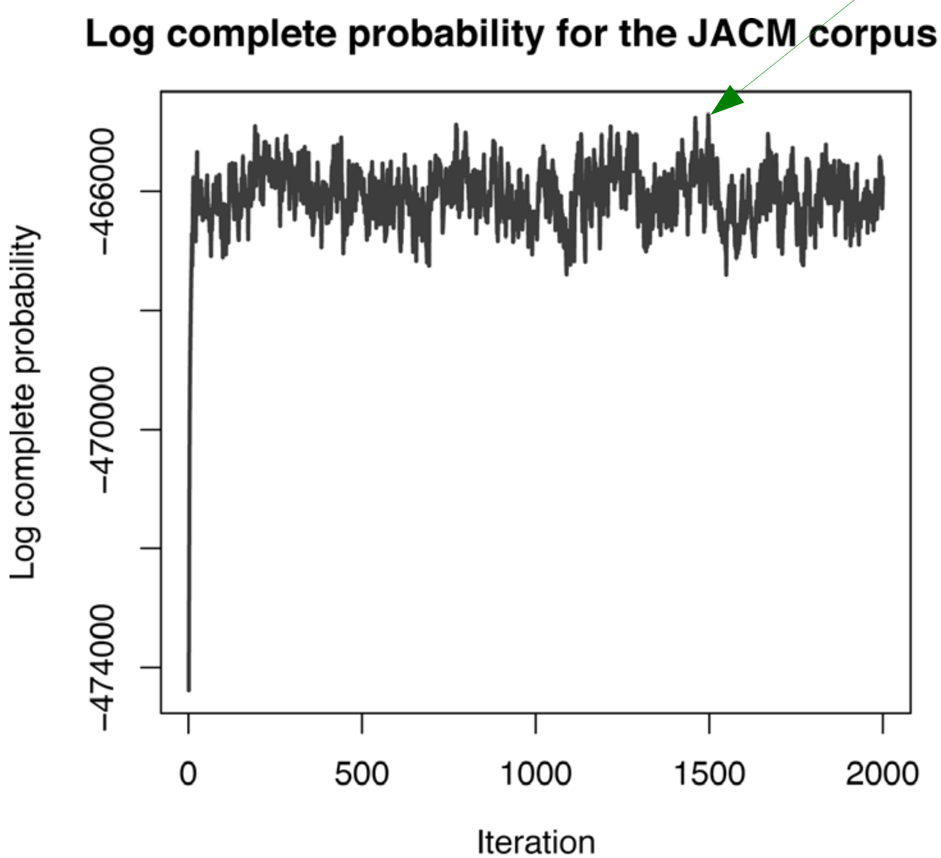
Put Metropolis-Hastings steps between iterations of the Gibbs sampler.

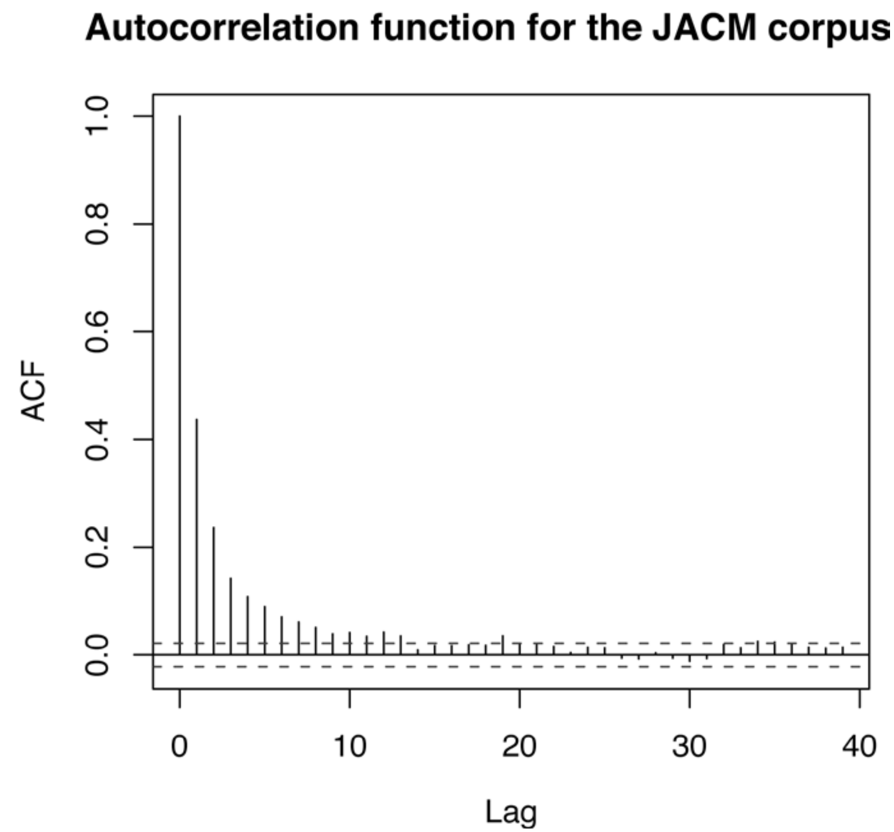
What purpose does this serve?
Resulting inference is less influenced by hyper-hyperparameters than by hyperparameters.

Convergence

$$\mathcal{L}^{(t)} = \log p\left(\mathbf{c}_{1:D}^{(t)}, \mathbf{z}_{1:D}^{(t)}, \mathbf{w}_{1:D} | \gamma, \eta, m, \pi\right)$$

Approximation of the posterior mode for these iterations

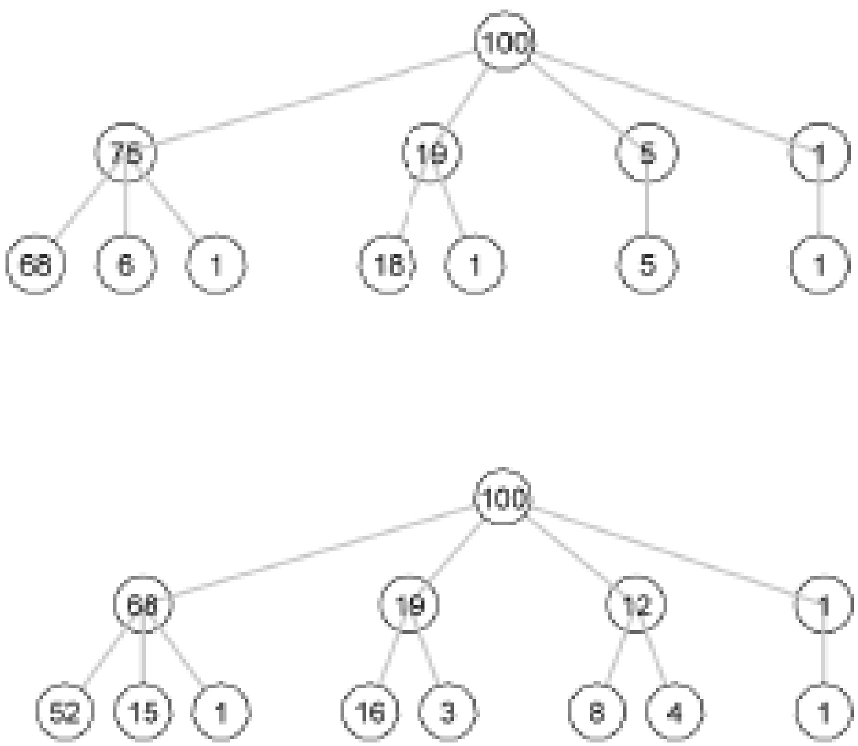




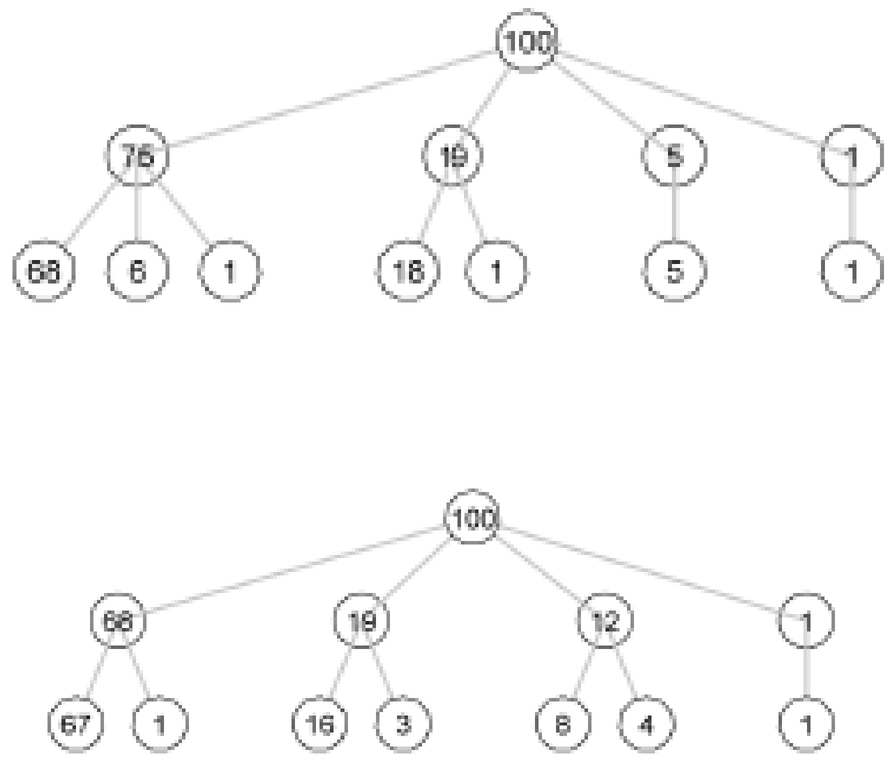
Experiments (simulated data)

- 100 documents drawn from an hLDA model.
- $\eta = 0.005; \quad \gamma = 1; \quad V = 100$

True dataset hierarchy



Posterior mode



Experiments (scientific abstracts)

- Fix parameters:

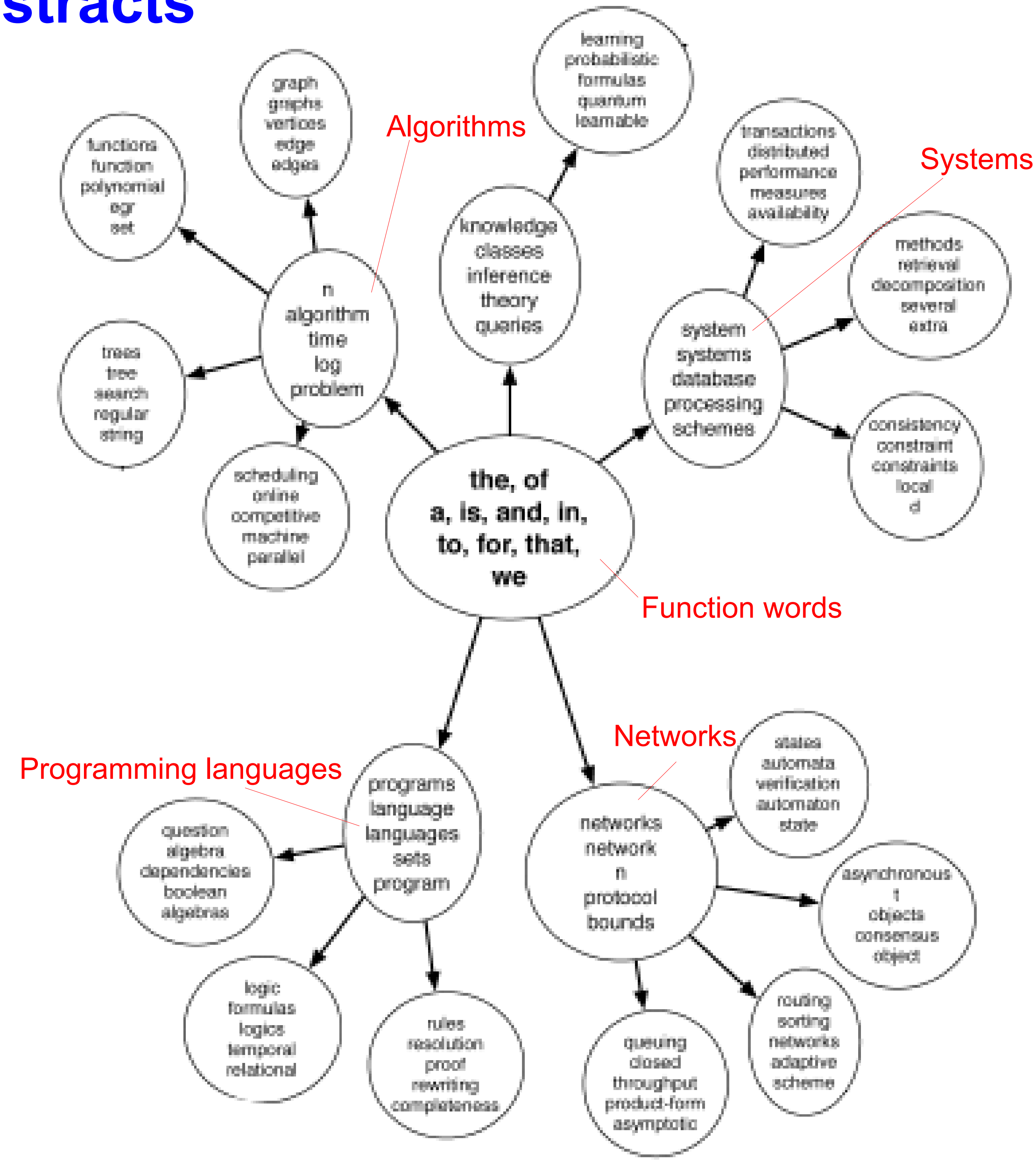
$$\eta = \{2.0, 1.0, 0.5\}; \quad \gamma = 1.0; \quad \pi = 100; \quad m = 0.5$$

- Recall that the topics have been integrated out, so they'll have to be estimated in the posterior. Posterior inference only yields a tree structure **c**, and assignments to levels **z**.

- The probability of a particular word *w* at a particular level *l* in a particular path **p** is:

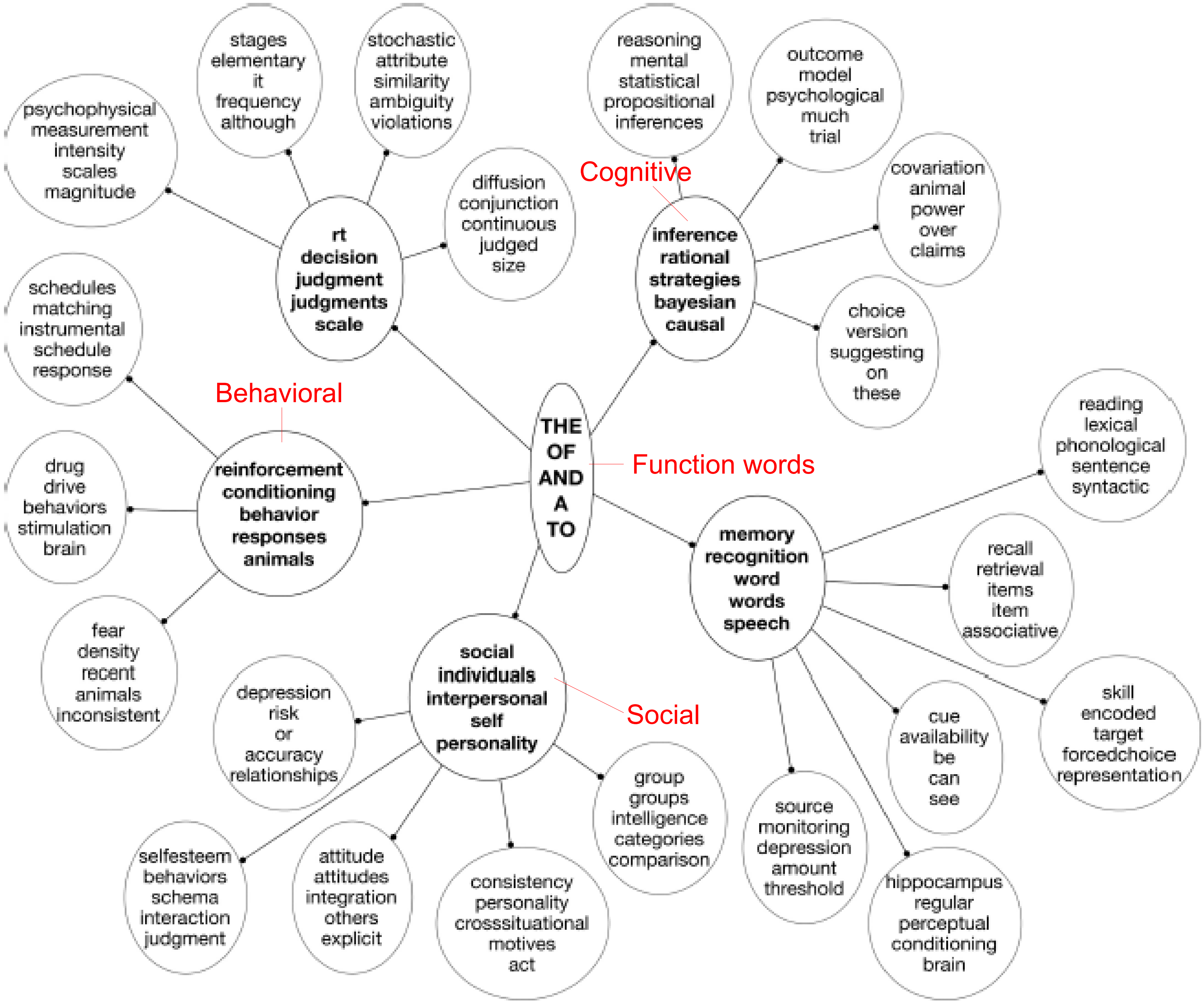
$$p(w|\mathbf{z}, \mathbf{c}, \mathbf{w}, \eta) = \frac{\#[\mathbf{z} = \ell, \mathbf{c} = \mathbf{p}, \mathbf{w} = w] + \eta}{\#[\mathbf{z} = \ell, \mathbf{c} = \mathbf{p}] + V\eta}$$

JACM abstracts

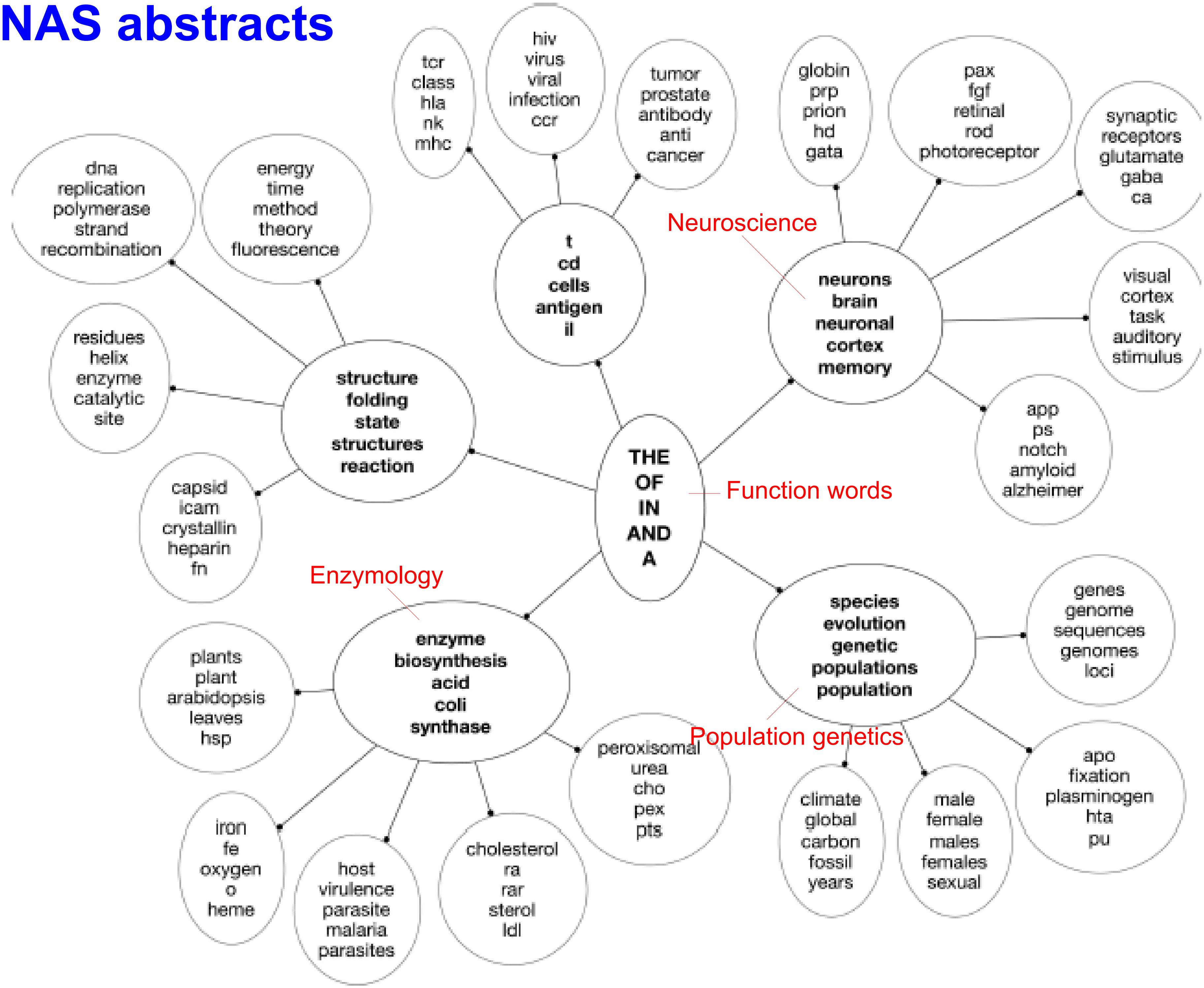


Slide 27

Psychological Review abstracts

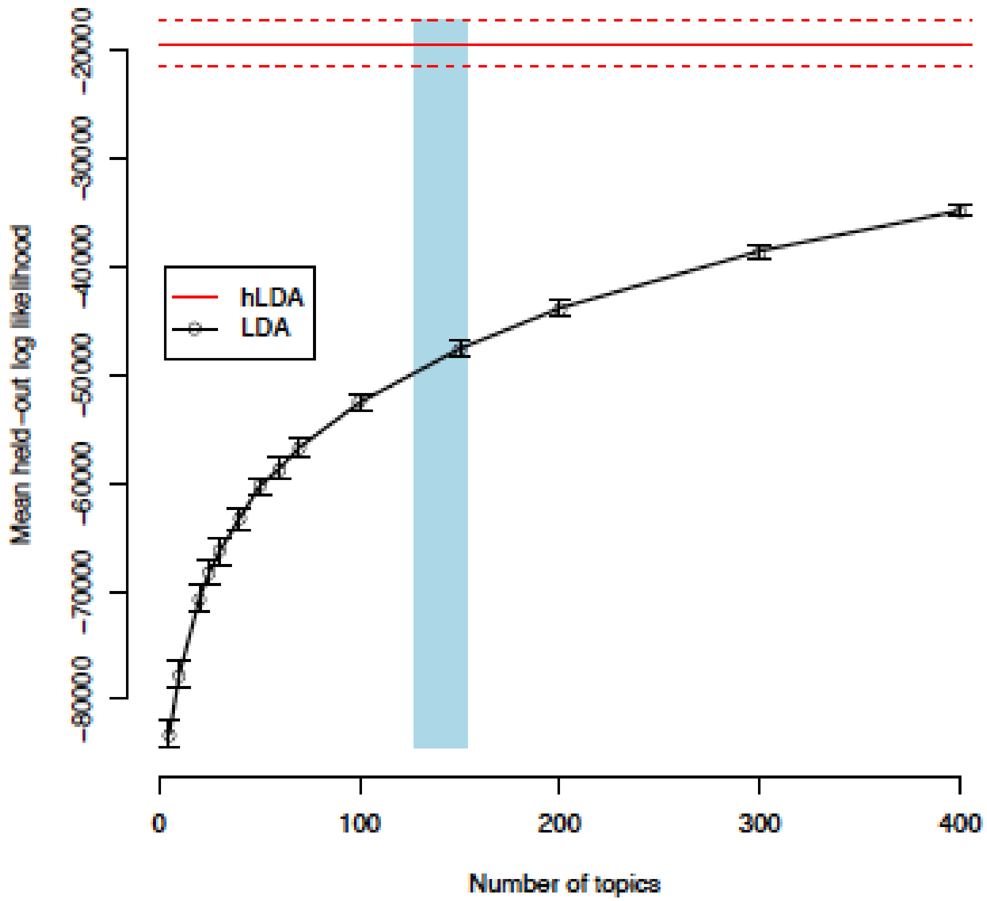


PNAS abstracts

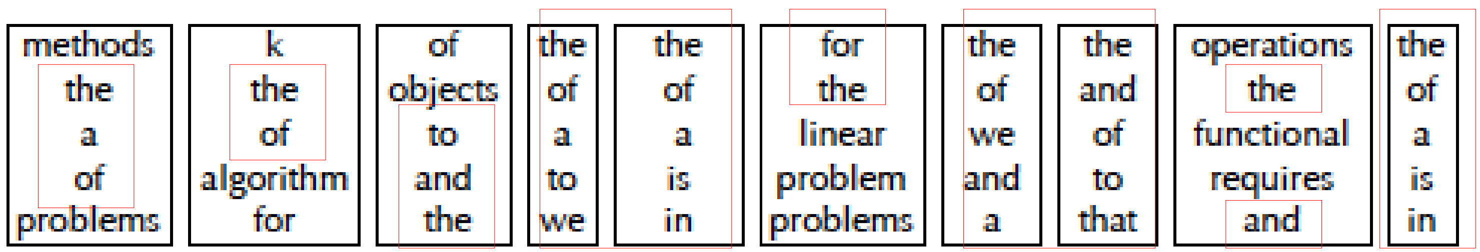


LDA v. hLDA

- Number of topics in LDA fixed beforehand
- In LDA, each document can place an arbitrary distribution over topics, not only those that lie on a path in the hierarchy.



LDA v. hLDA



Function words everywhere

Conclusions

- **The Good:**
  - The nested CRP allows a flexible family of prior distributions over arbitrary tree structures; definitely could be useful for more than just topic models.
  - Nice qualitative results for topic hierarchies.
  - Same inference of number of topics as a model like HDP
- **The Bad/Ugly:**
  - The restriction that documents can only follow a single path in the tree is a possibly limiting one. (Michael Jordan talked about this extension when he spoke in class).
  - Quantitative evaluation is not extensive enough.
  - I'd like to see comparisons of hLDA with HDP, as opposed to LDA. It seems like that would get closer to the heart of whether hierarchies are helpful or not.