Pachinko Allocation

2 Papers

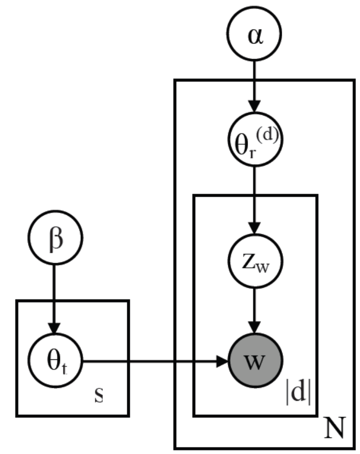
Presented by Evan Ribnick

Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations

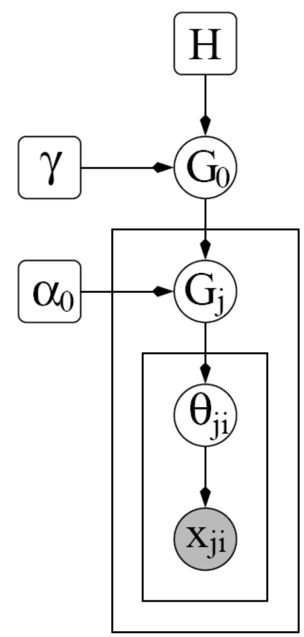
Wei Li and Andrew McCallum

Motivation

- LDA
  - Does not model correlations among topics

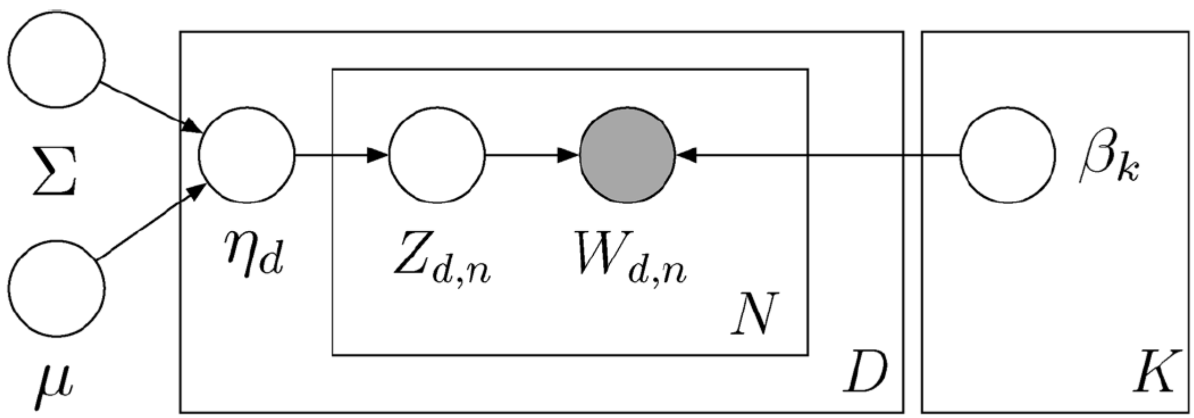


- HDP
  - Topic correlations from base measures of Dirichlet prior



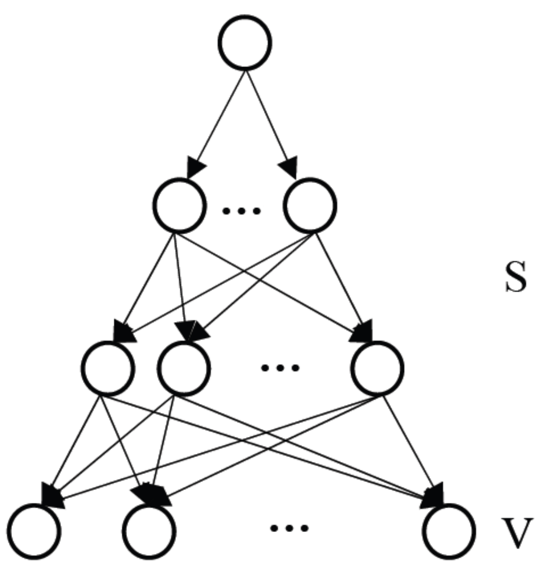
Motivation (cont'd)

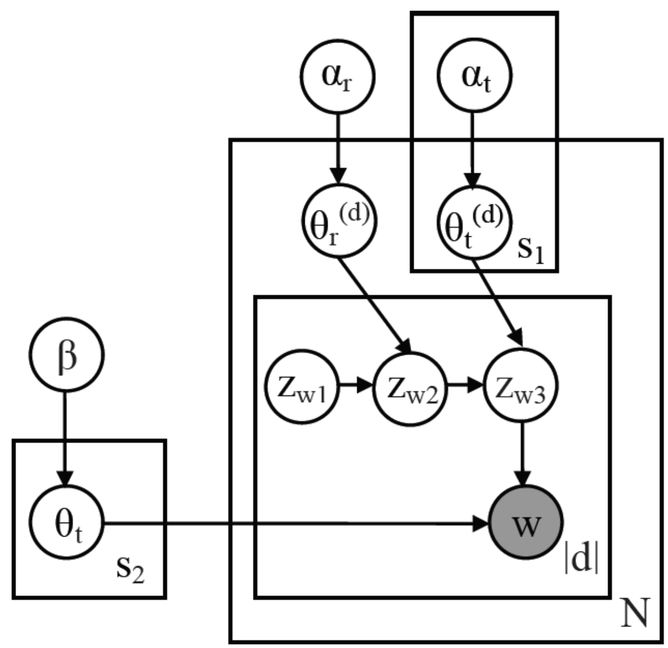
- CTM
  - Mixture weights: logistic normal, pairwise topic correlations



PAM

- Pachinko allocation model (PAM)
- Pachinko: Japanese game, balls fall through pins from top to bottom
- Explicitly represent arbitrary topic correlations





The Model

- Topics, sub-topics
  - Each topic is a Dirichlet distribution

- Generative model
  1. Sample a multinomial from each topic's Dirichlet
  2. Starting from top of tree, sample from multinomials, moving down tree
  3. At bottom, sample from sub-topic multinomial for a word

- This paper: only 4-level PAM

The Model (cont'd)

- Joint prob. of document, path, multinomials:

$$P(d, \mathbf{z}^{(d)}, \theta^{(d)}|\alpha) = \prod_{i=1}^{s} P(\theta_{t_i}^{(d)}|\alpha_i) \times \prod_{w} (\prod_{i=2}^{L_w} P(z_{wi}|\theta_{z_{w(i-1)}}^{(d)}) P(w|\theta_{z_{wL_w}}^{(d)}))$$

- Marginal:

$$P(d|\alpha) = \int \prod_{i=1}^{s} P(\theta_{t_i}^{(d)}|\alpha_i) \times \prod_{w} \sum_{\mathbf{z}_w} (\prod_{i=2}^{L_w} P(z_{wi}|\theta_{z_{w(i-1)}}^{(d)}) P(w|\theta_{z_{wL_w}}^{(d)})) \mathrm{d}\theta^{(d)}$$

Inference

- Gibbs sampling
- For each word
  - Sample a topic path, enumerating all possibilities

$$P(z_{w2} = t_k, z_{w3} = t_p | \mathbf{D}, \mathbf{z}_{-w}, \alpha, \beta) \propto$$
$$\frac{n_{1k}^{(d)} + \alpha_{1k}}{n_1^{(d)} + \sum_{k'} \alpha_{1k'}} \times \frac{n_{kp}^{(d)} + \alpha_{kp}}{n_k^{(d)} + \sum_{p'} \alpha_{kp'}} \times \frac{n_{pw} + \beta_w}{n_p + \sum_m \beta_m}$$

  - n: empirical frequencies
  - alpha: parameter of root and super-topic Dirichlet
  - Beta: parameter of sub-topic Dirichlet

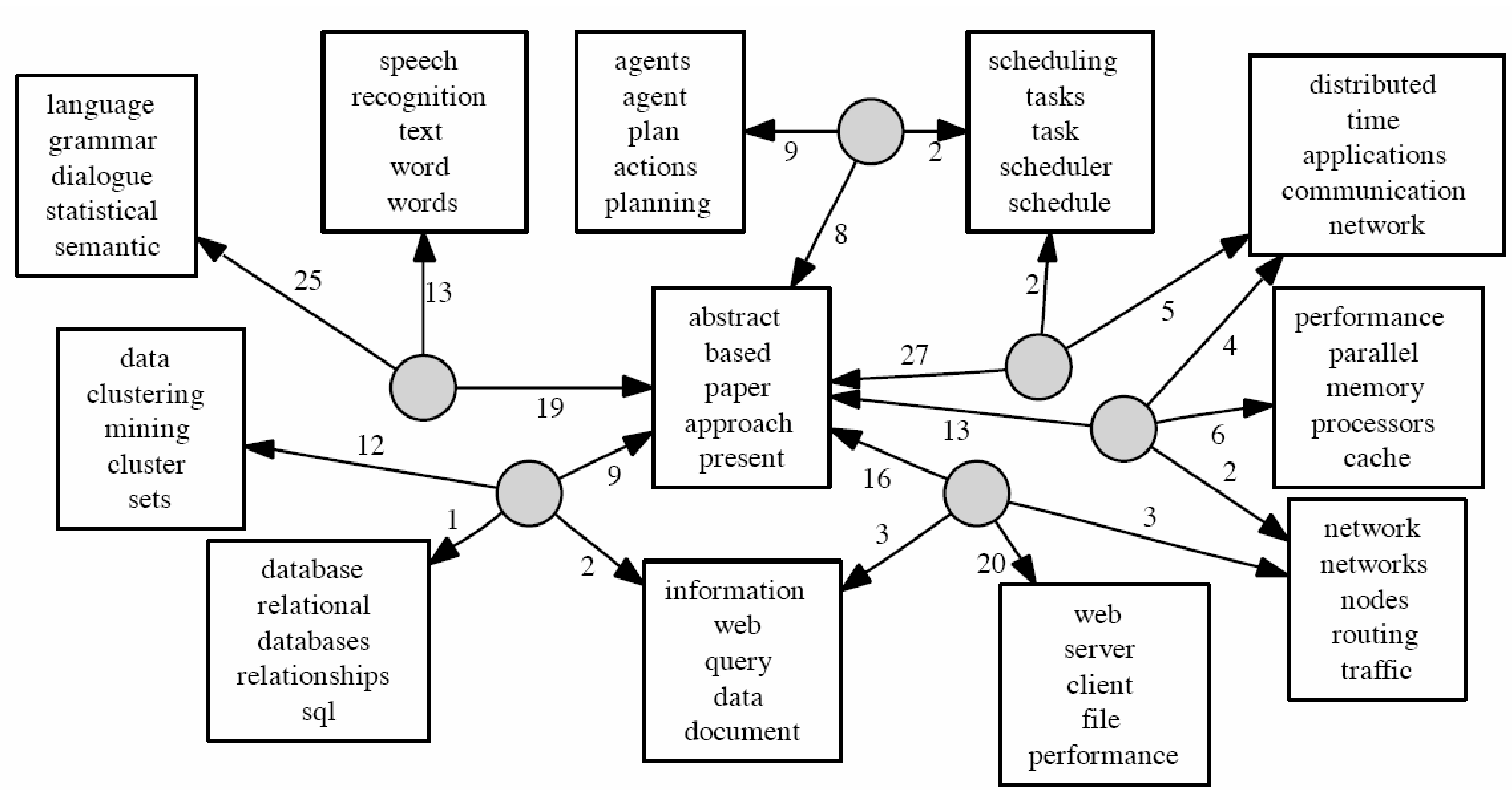
Parameter Estimation

- Need to estimate Dirichlet parameters *alpha* for super-topics
- At each iteration of sampling:

$$
\begin{aligned}
mean_{xy} &= \frac{1}{N} \times \sum_d \frac{n_{xy}^{(d)}}{n_x^{(d)}}; \\
var_{xy} &= \frac{1}{N} \times \sum_d (\frac{n_{xy}^{(d)}}{n_x^{(d)}} - mean_{xy})^2; \\
m_{xy} &= \frac{mean_{xy} \times (1 - mean_{xy})}{var_{xy}} - 1; \\
\alpha_{xy} &\propto mean_{xy}; \\
\sum_y \alpha_{xy} &= \frac{1}{5} \times exp(\frac{\sum_y log(m_{xy})}{s_2 - 1}).
\end{aligned}
$$

Results

- Rexa



Results (cont'd)

- Human judgment
  - Which topic description has stronger sense of semantic coherence and specificity?

PAM	LDA	PAM	LDA
control	control	motion	image
systems	systems	image	motion
robot	based	detection	images
adaptive	adaptive	images	multiple
environment	direct	scene	local
goal	con	vision	generated
state	controller	texture	noisy
controller	change	segmentation	optical
5 votes	0 vote	4 votes	1 vote

PAM	LDA	PAM	LDA
signals	signal	algorithm	algorithm
source	signals	learning	algorithms
separation	single	algorithms	gradient
eeg	time	gradient	convergence
sources	low	convergence	stochastic
blind	source	function	line
single	temporal	stochastic	descent
event	processing	weight	converge
4 votes	1 vote	1 vote	4 votes

	LDA	PAM
5 votes	0	5
$\geq$ 4 votes	3	8
$\geq$ 3 votes	9	16

Results (cont'd)

- Likelihood comparison on holdout set
  - NIPS data
  - PAM and LDA the best
  - PAM better for larger number of topics
- Document classification accuracy

class	# docs	LDA	PAM
graphics	243	83.95	86.83
os	239	81.59	84.10
pc	245	83.67	88.16
mac	239	86.61	89.54
windows.x	243	88.07	92.20
total	1209	84.70	87.34

Conclusion

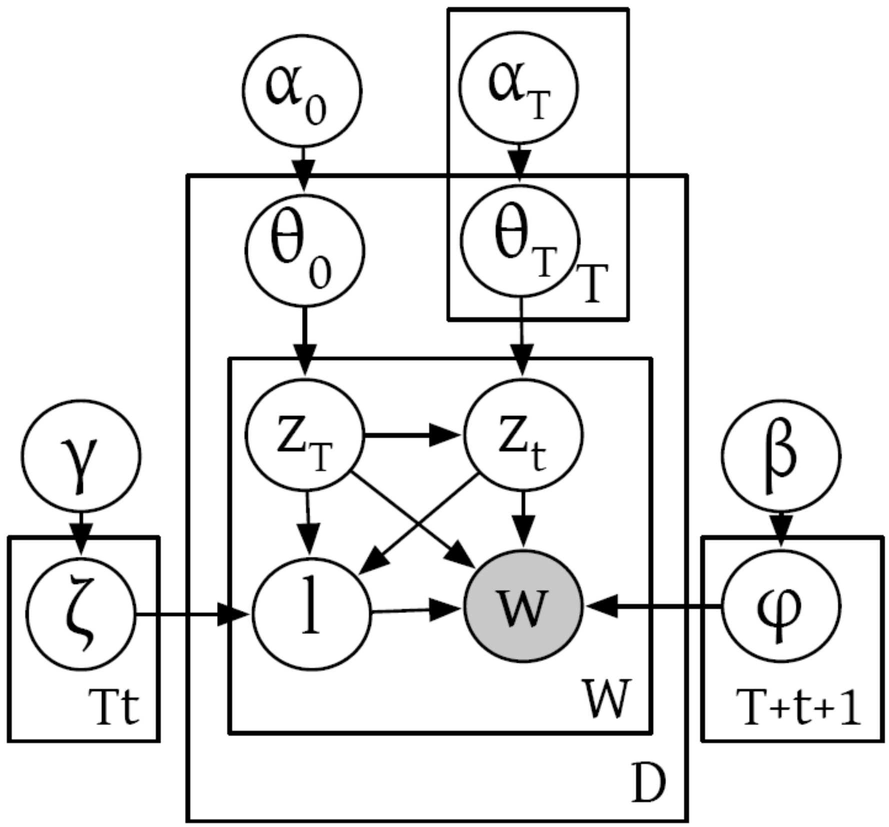
- Main contribution: a model which captures correlations between topics
- Model is flexible
  - Could use any distribution in nodes
- Problem is well motivated

Mixtures of Hierarchical Topics with Pachinko Allocation

David Mimmo, Wei Li and Andrew McCallum

hPAM

- Extension of PAM
- *Every* node has distribution over words



hPAM1

- Generative model:

1. For each document d, sample a distribution θ_0 over super-topics and a distribution θ_T over sub-topics for each super-topic.
2. For each word w,
   (a) Sample a super-topic z_T from θ_0.
   (b) Sample a sub-topic z_t from θ_{z_T}.
   (c) Sample a level ℓ from $\zeta_{z_T z_t}$.
   (d) Sample a word from ϕ_0 if $\ell = 1$, ϕ_{z_T} if $\ell = 2$, or ϕ_{z_t} if $\ell = 3$.

hPAM2

- Generative model:

1. For each document d, sample a distribution θ_0 over super-topics and a distribution θ_T over sub-topics for each super-topic.
2. For each word w,
   (a) Sample a super-topic z_T from θ_0. If $z_T = 0$, sample a word from ϕ_0.
   (b) Otherwise, sample a sub-topic z_t from θ_{z_T}. If $z_t = 0$, sample a word from ϕ_{z_T}.
   (c) Otherwise, sample a word from ϕ_{z_t}.

Inference

- Gibbs sampling
- For each word
  - Sample super-topic, sub-topic, and level
  - For hPAM1:

$$p(z_{Ti}, z_{ti}, \ell_i | \mathbf{w}, \mathbf{z}_{\mathbf{T}\setminus\mathbf{i}}, \mathbf{z}_{\mathbf{t}\setminus\mathbf{i}}, \ell_{\setminus\mathbf{i}}, \alpha, \beta, \gamma) \quad \propto$$

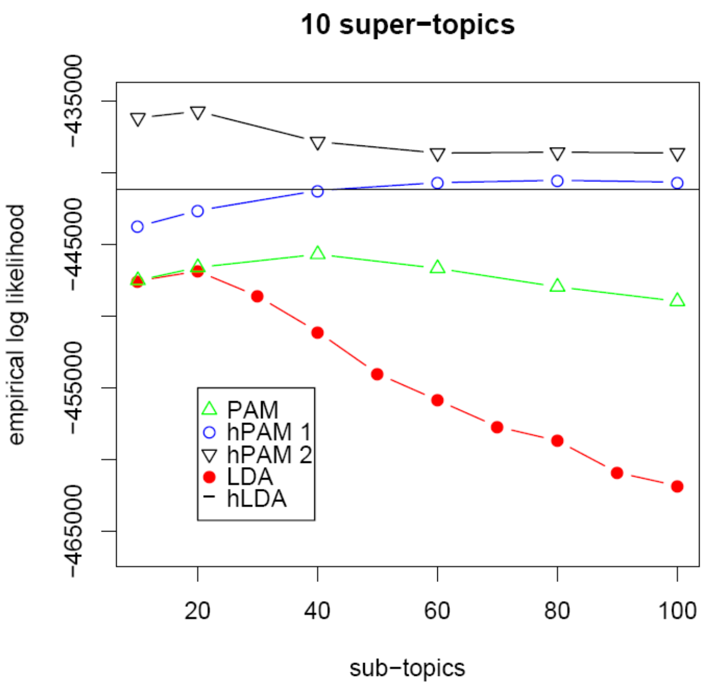
$$\frac{\alpha_T + N_d^T}{\sum_{T'} \alpha_{T'} + N_d} \frac{\alpha_{Tt} + N_d^{Tt}}{\sum_{t'} \alpha_{Tt'} + N_d^T} \times$$

$$\frac{\gamma + N_{Tt}^{\ell}}{3\gamma + N_{Tt}} \frac{\beta_w + N_{Tt\ell}^w}{\sum_{w'} \beta_{w'} + N_{Tt\ell}}.$$

- Speed up: marginal distr. over output topics

Results

- Likelihood on holdout set: Medline DB
  - Train model on training set
  - Calculate empirical distribution over words drawn from model
  - Calculate likelihood of holdout documents
  - Repeated for different numbers of super-topics
  - hPAM better for larger number of topics (finer granularity)



Results (cont'd)

- hPAM1 combines high topic/journal MI and high empirical log-likelihood
- Quality of topics: qualitative

- virus infection cells infected cell viral gene replication
  - rna replication virus dna viral
    - results study specific studies role
    - protein proteins binding virus domain
    - gene genes expression sequence protein
  - spinal nerve pain cord rats
    - rats receptor induced kg administration
    - ca neurons glutamate receptor hippocampal
    - neurons nucleus expression cells fos
  - cardiac heart ventricular myocardial left
    - patients risk years clinical ci
    - disease risk ad women subjects
    - levels increased significantly compared
  - cells cd cell marrow specific
    - results study specific studies role
    - leukemia cell expression aml myeloid
    - patients therapy treatment disease dose

Conclusion

- Main contribution: model hierarchical structure of topics and their interdependencies
- Relatively simple extension of PAM
- Could use other configurations
  - Not all subtopics must be shared . . .