Reliable Variational Learning for Hierarchical Dirichlet Processes

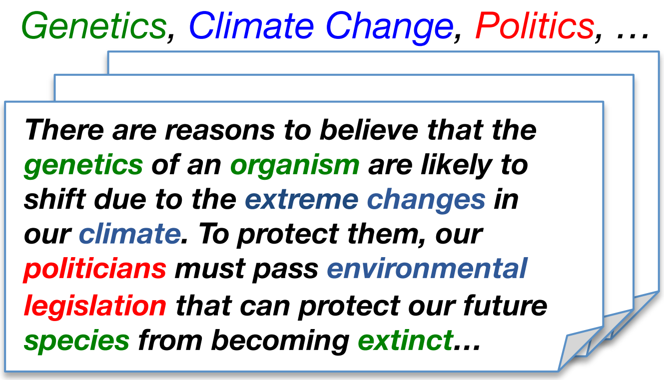
Erik Sudderth

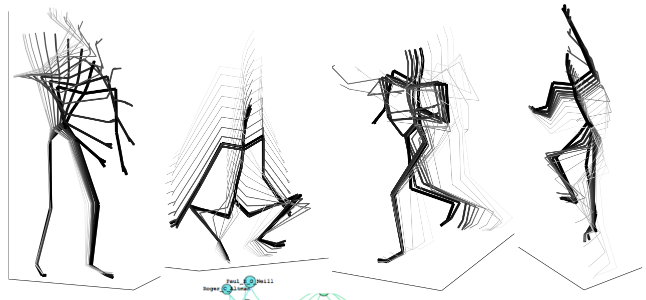
Brown University Computer Science

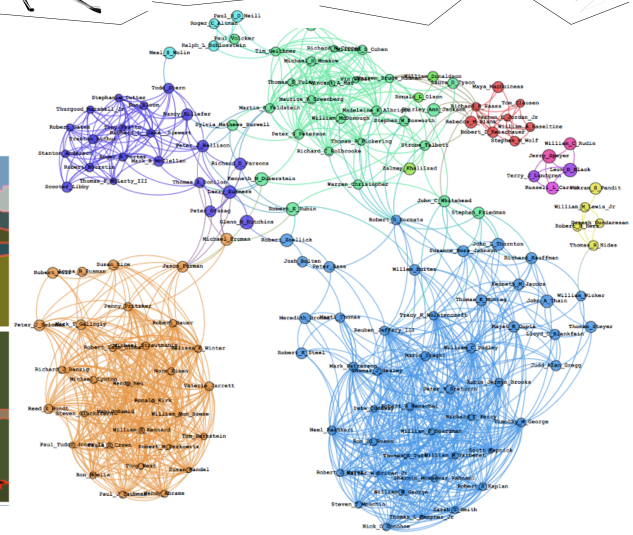
Collaborators:

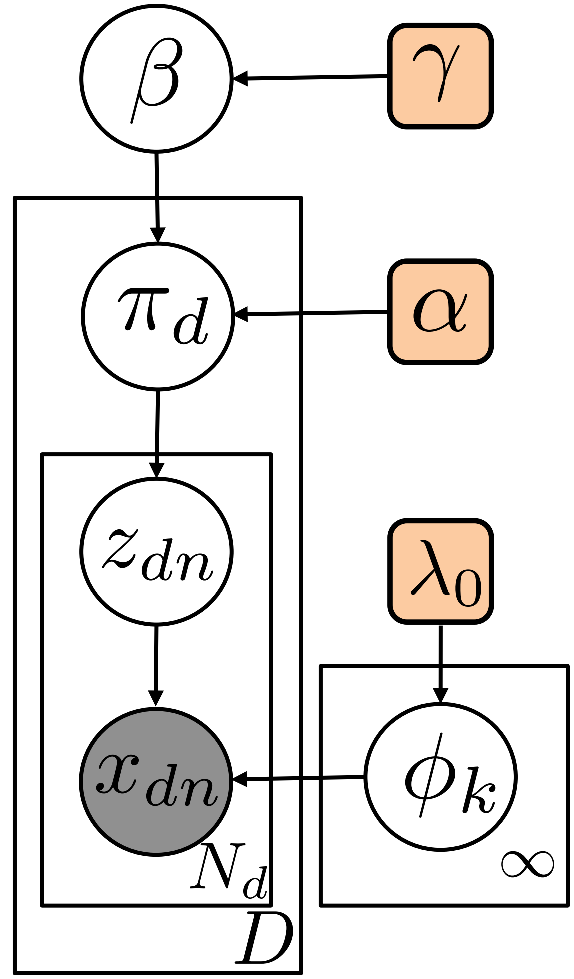
- *Michael Hughes & Dae Il Kim, Brown University*
- *Prem Gopalan & David Blei, Princeton University*

Learning Structured BNP Models









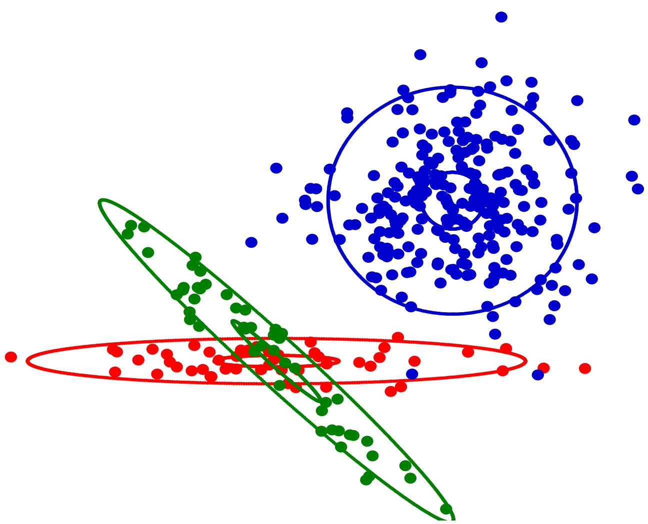
- ***Nonparametric:*** Data-driven discovery of model structure: *topics, behaviors, objects, communities…*
- ***Reliable:*** Structure driven by data and modeling assumptions, not heuristic algorithm initializations
- ***Parsimonious:*** Want a single model structure with good predictive power, not full posterior uncertainty

Hierarchical Dirichlet Process
(Teh et al., JASA 2006)

Memoized Variational Inference for Dirichlet Process Mixture Models

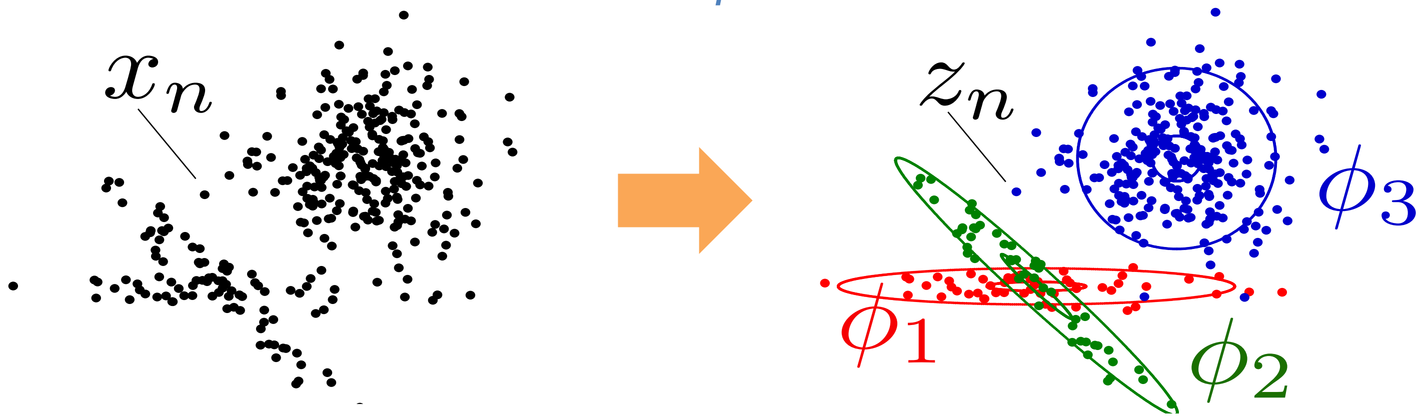
Michael Hughes & E. Sudderth

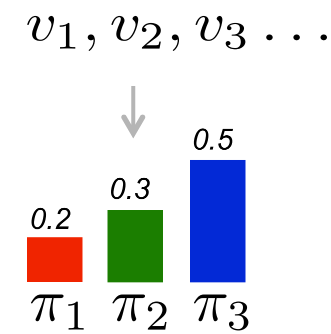
2013 Conference on Neural Information Processing Systems



Dirichlet Process Stick-Breaking

GOAL: *Partition data into an a priori unknown number of discrete clusters.*





$$\pi \sim \mathrm{Stick}(\alpha)$$

Each cluster k = 1, 2, …

- Cluster shape: $\phi_k \sim H(\lambda_0)$
- Stick proportion: $v_k \sim \mathrm{Beta}(1, \alpha)$
- Cluster frequency: $\pi_k = v_k \prod_{\ell=1}^{k-1}(1 - v_\ell)$

Stick-Breaking
(Sethuraman 1994)

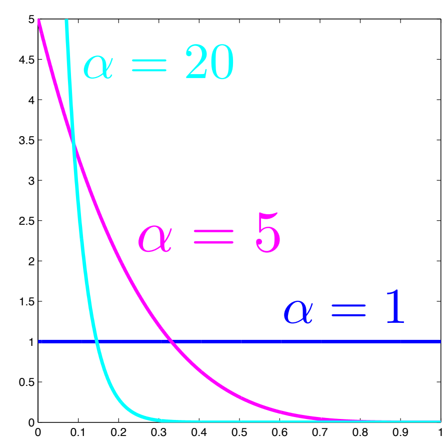
$\pi_1 = v_1$

0 — 1

$\pi_2 = v_2(1 - v_1)$

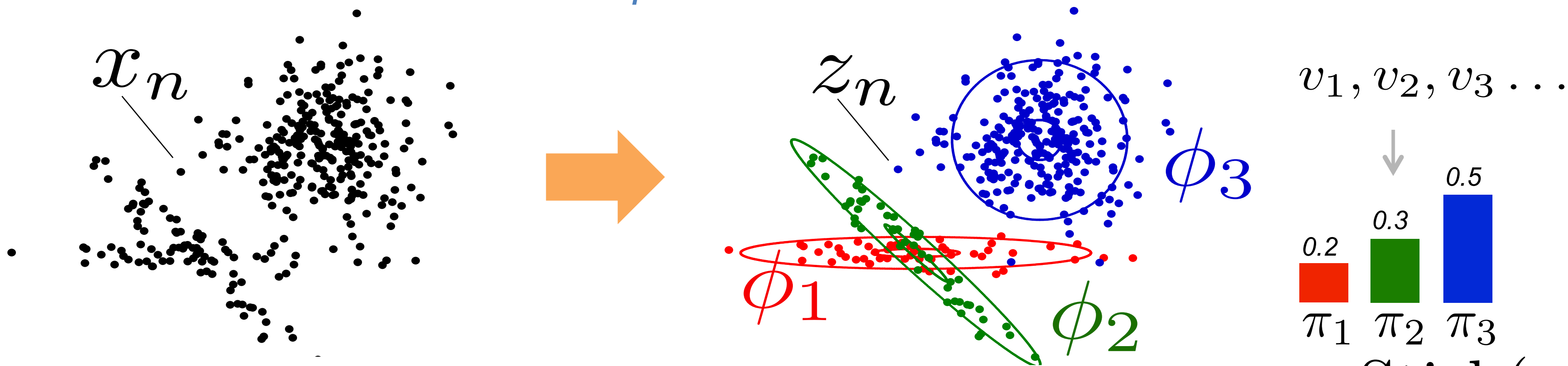
$\pi_3 = v_3(1 - v_2)(1 - v_1)$

$$1 - \sum_{k=1}^{K} \pi_k = \prod_{k=1}^{K}(1 - v_k)$$



Dirichlet Process Mixtures

***GOAL:** Partition data into an a priori unknown number of discrete clusters.*



$$\pi \sim \mathrm{Stick}(\alpha)$$

Each cluster k = 1, 2, …

- Cluster shape: $\phi_k \sim H(\lambda_0)$
- Stick proportion: $v_k \sim \mathrm{Beta}(1, \alpha)$
- Cluster frequency: π_k

Each observation n = 1, 2, …, N:

- Cluster assignment: $z_n \sim \mathrm{Cat}(\pi)$
- Observed value: $x_n \sim \mathrm{F}(\phi_{z_n})$

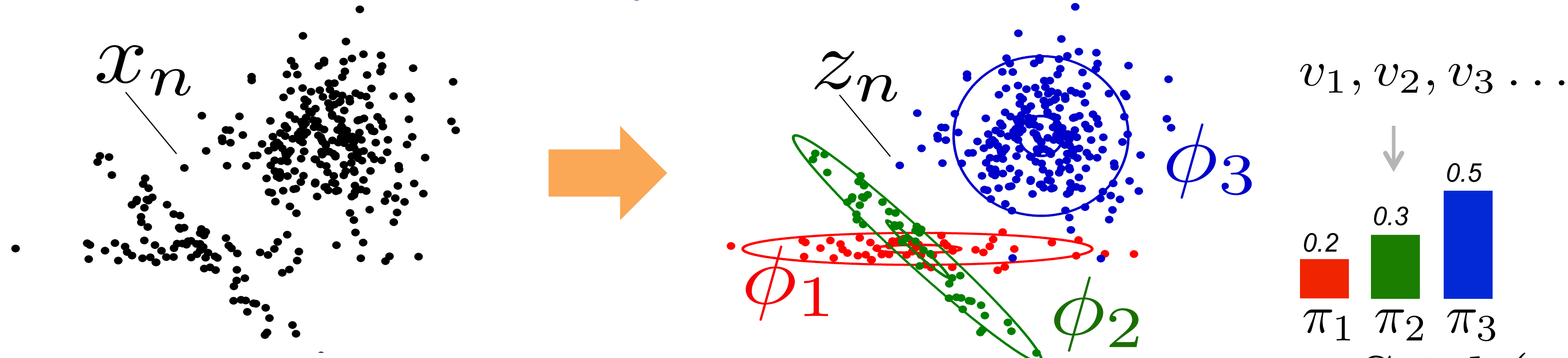
Assume exponential family likelihoods with conjugate priors

$$f(x_n \mid \phi_k) = \exp\big(\phi_k^T t(x_n) - a(\phi_k)\big)$$

$$h(\phi_k \mid \lambda_0) = \exp\big(\lambda_0^T \bar{t}(\phi_k) - \bar{a}(\lambda_0)\big), \quad \bar{t}(\phi_k) = [\phi_k, -a(\phi_k)]$$

Dirichlet Process Mixtures

GOAL: *Partition data into an a priori unknown number of discrete clusters.*



$$\pi \sim \mathrm{Stick}(\alpha)$$

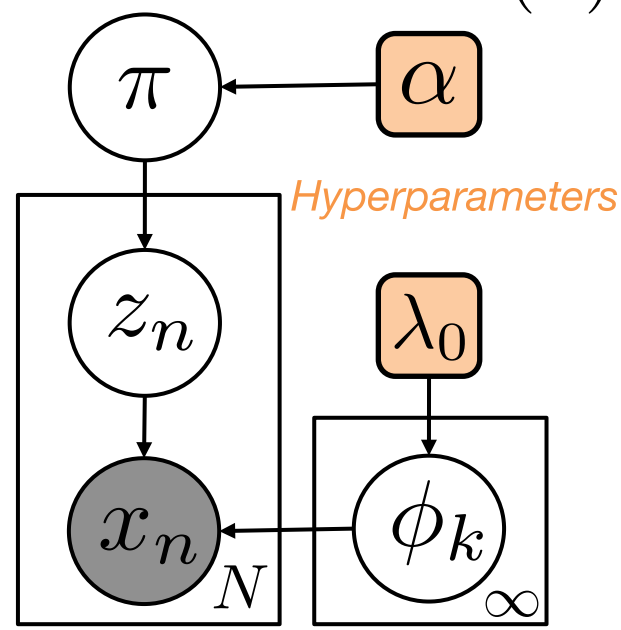
Each cluster k = 1, 2, …

- Cluster shape: $\phi_k \sim H(\lambda_0)$
- Stick proportion: $v_k \sim \mathrm{Beta}(1, \alpha)$
- Cluster frequency: π_k

Each observation n = 1, 2, …, N:

- Cluster assignment: $z_n \sim \mathrm{Cat}(\pi)$
- Observed value: $x_n \sim \mathrm{F}(\phi_{z_n})$

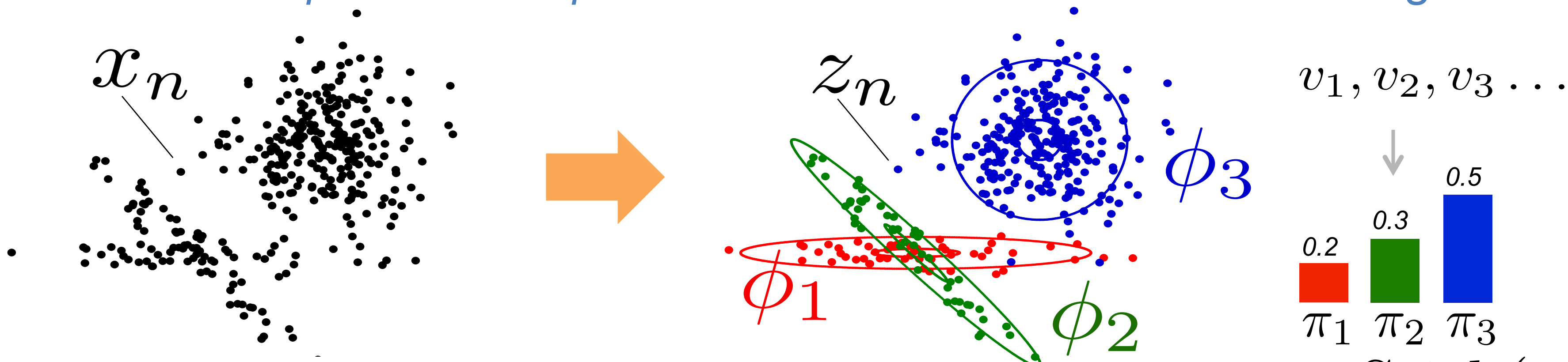
$$f(x_n \mid z_n = k, \phi) = \exp\left(\phi_k^T t(x_n) - a(\phi_k)\right)$$



Visually summarize model structure via directed graphical model

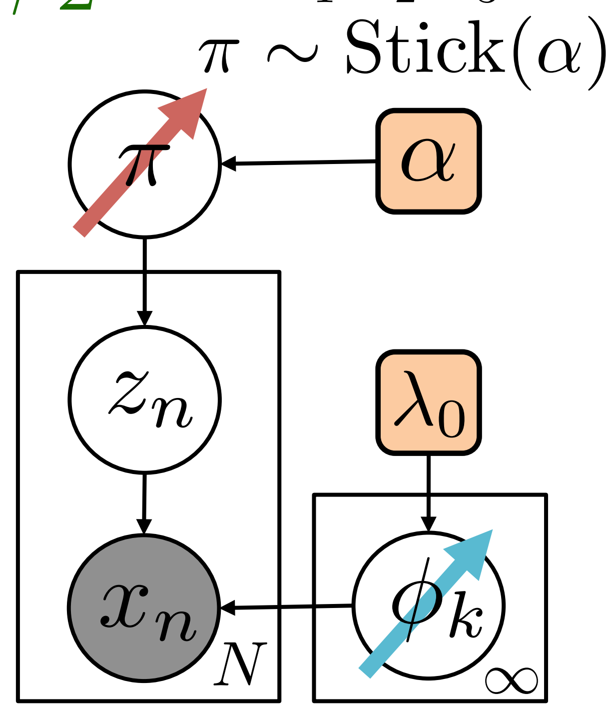
MCMC for DP Mixtures

Can we sample from the posterior distribution over data clusterings?



Given any fixed partition z:

- Marginalize stick-breaking weights via *Chinese Restaurant Process*, assigning positive probability to all partitions of data (large support)
- Via *conjugacy* of base measure to exponential family likelihood, marginalize cluster shape parameters

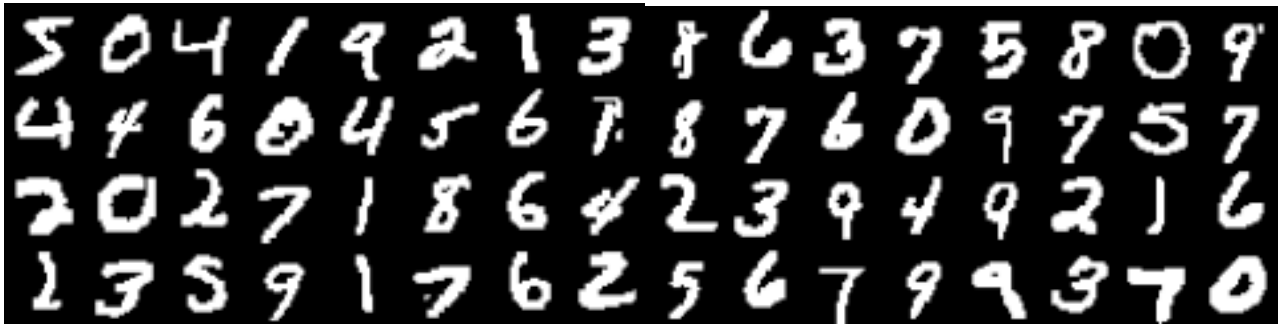


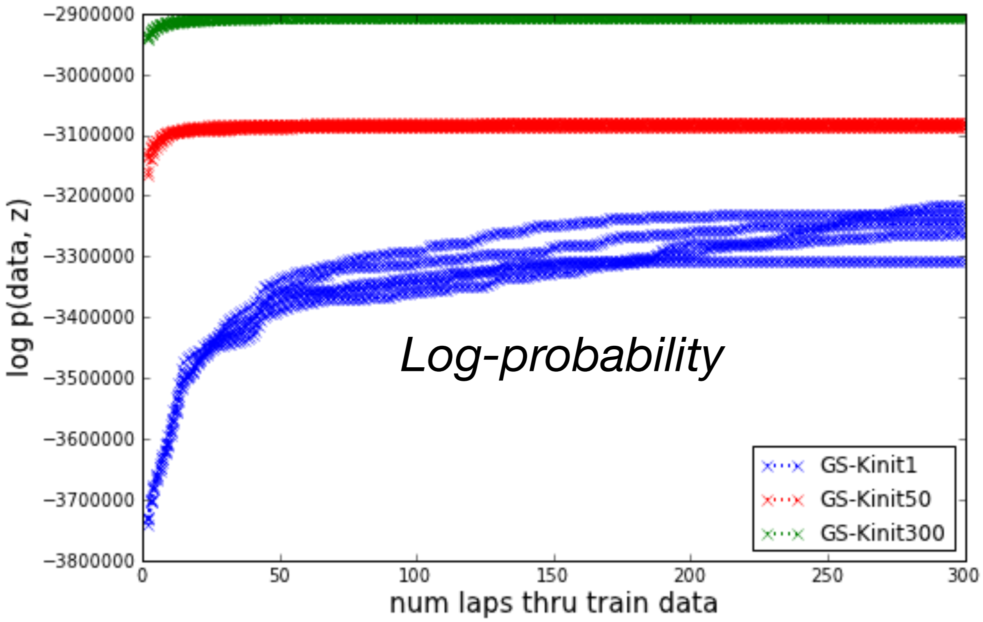
Gibbs Sampler: *(Neal 1992, MacEachern 1994)*

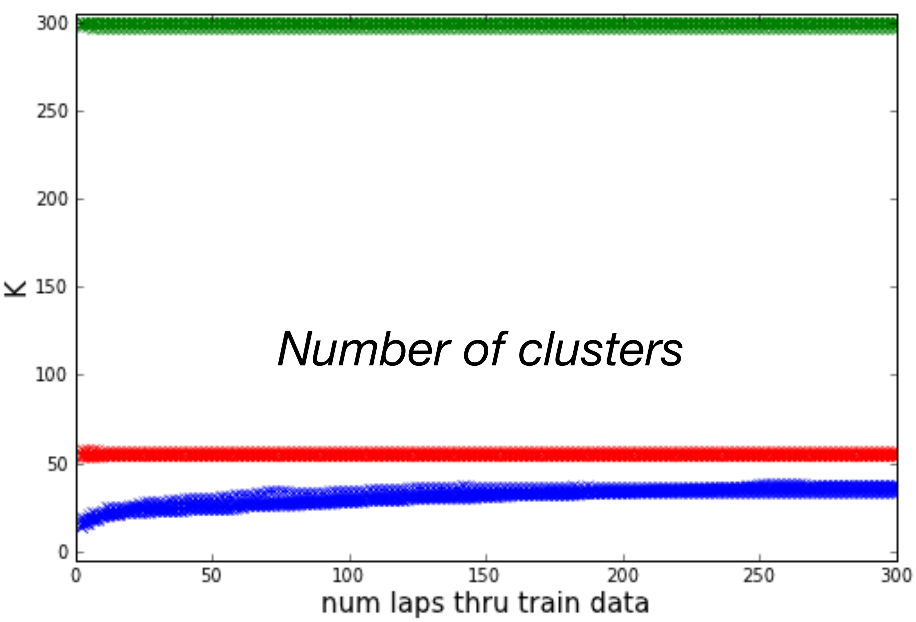
Iteratively resample cluster assignment for one observation, fixing all others.

Mixing for DP Mixture Samplers

MNIST: *60,000 digits projected to 50 dimensions via PCA.*







- Five random initializations from *K=1*, *K=50*, *K=300* clusters
- Reversible jump MCMC? Proposals slow, acceptance low.

Variational Bounds

What is the marginal likelihood of our observed data?

$$\log p(x \mid \alpha, \lambda_0) = \log \sum_z \iint p(x, z, v, \phi \mid \alpha, \lambda_0)\, dv d\phi$$

$$= \log \sum_z \iint \frac{q(z, v, \phi) p(x, z, v, \phi \mid \alpha, \lambda_0)}{q(z, v, \phi)}\, dv d\phi$$

$$= \log \mathbb{E}_q \left[\frac{p(x, z, v, \phi \mid \alpha, \lambda_0)}{q(z, v, \phi)} \right]$$

Expectation with respect to some variational distribution $q(z, v, \phi)$

Jensen's Inequality

$$\geq \mathbb{E}_q[\log p(x, z, v, \phi \mid \alpha, \lambda_0)] - \mathbb{E}_q[\log q(z, v, \phi)] = \mathcal{L}(q)$$

Expected log-likelihood (negative of "average energy")

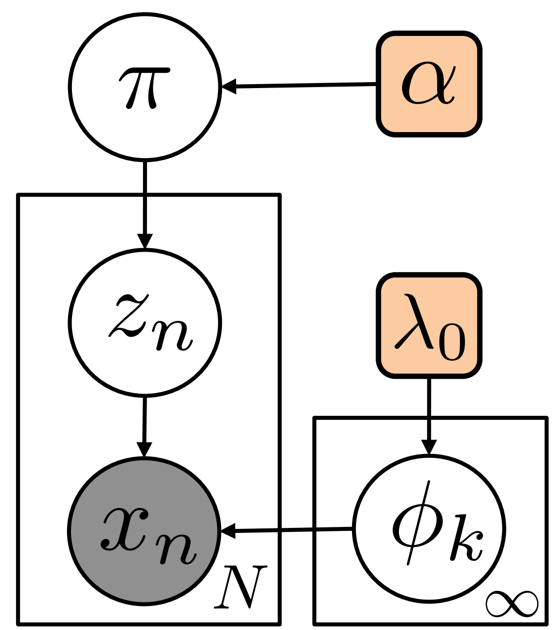
Variational entropy

- Maximizing this bound recovers true posterior:

$$\mathcal{L}(q) = \log p(x \mid \alpha, \lambda_0) - \mathrm{KL}(q(z, v, \phi) \,||\, p(z, v, \phi \mid x, \alpha, \lambda_0))$$

- The simplest *mean field* variational methods create tractable algorithms via *assumed independence*:

$$q(z, v, \phi) = q(z) q(v, \phi)$$



Approximating Infinite Models

$$q(z_n = k) = r_{nk}$$

Beta Distribution

Exponential Family from Conjugate Prior

$$q(z, v, \phi) = q(z)q(v, \phi) = \left[\prod_{n=1}^{N} q(z_n)\right] \cdot \left[\prod_{k=1}^{\infty} q(v_k)q(\phi_k)\right]$$

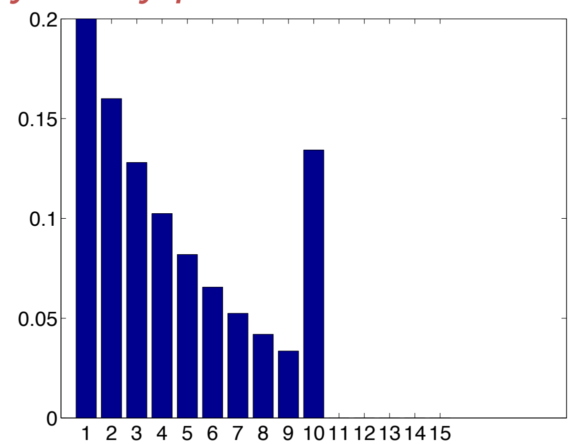
Categorical distribution with unbounded support, and infinitely many potential clusters!

Top-Down Model Truncation

Blei & Jordan, 2006; Ishwaran & James, 2001

$$q(z_n) = \mathrm{Cat}(z_n \mid r_{n1}, r_{n2}, \ldots, r_{nK})$$

$$q(v, \phi) = \left[\prod_{k=1}^{K} q(\phi_k)\right] \cdot \left[\prod_{k=1}^{K-1} q(v_k)\right], \quad v_K = \prod_{k=1}^{K-1} (1 - v_k).$$



$\alpha = 4, K = 10$

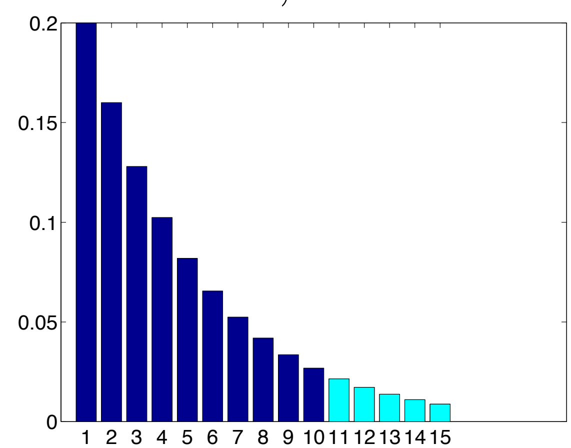
Bottom-Up Assignment Truncation

Bryant & Sudderth, 2012; Teh, Kurihara, & Welling, 2008

$$q(z_n) = \mathrm{Cat}(z_n \mid r_{n1}, r_{n2}, \ldots, r_{nK}, 0, 0, 0, \ldots)$$

$$q(v, \phi) = \prod_{k=1}^{\infty} q(v_k)q(\phi_k)$$

For any k>K, optimal variational distributions equal prior & need not be explicitly represented



Batch Variational Updates

A Bayesian nonparametric analog of Expectation-Maximization (EM)

$$q(z,v,\phi) = \left[\prod_{n=1}^{N} q(z_n \mid r_n)\right] \cdot \left[\prod_{k=1}^{\infty} \mathrm{Beta}(v_k \mid \alpha_{k1}, \alpha_{k0}) h(\phi_k \mid \lambda_k)\right]$$

$$q(z_n) = \mathrm{Cat}(z_n \mid r_{n1}, r_{n2}, \ldots, r_{nK}, 0, 0, 0, \ldots)$$ *for some K>0*

Update Assignments (The Expectation Step): *For all N data,*

$$r_{nk} \propto \exp(\mathbb{E}_q[\log \pi_k(v)] + \mathbb{E}_q[\log p(x_n \mid \phi_k)]) \qquad \text{for } k \leq K$$

$$\mathbb{E}_q[\log \pi_k(v)] = \underbrace{\mathbb{E}_q[\log(v_k)]}_{\psi(\alpha_{k1}) - \psi(\alpha_{k1} + \alpha_{k0})} + \sum_{\ell=1}^{k-1} \underbrace{\mathbb{E}_q[\log(1 - v_\ell)]}_{\psi(\alpha_{k0}) - \psi(\alpha_{k1} + \alpha_{k0})}$$

Update Cluster Parameters (The Other Expectation Step):

$$N_k^0 = \sum_{n=1}^{N} r_{nk} \qquad s_k^0 \leftarrow \sum_{n=1}^{N} r_{nk} t(x_n) \qquad \lambda_k \leftarrow \lambda_0 + s_k^0$$

Expected counts and sufficient statistics are only non-zero for first K clusters

$$\alpha_{k1} \leftarrow 1 + N_k^0$$

$$\alpha_{k0} \leftarrow \alpha + \sum_{\ell=k+1}^{\infty} N_\ell^0 = \alpha + \sum_{\ell=k+1}^{K} N_\ell^0$$

$$\mathbb{E}_q[v_k] = \frac{\alpha_{k1}}{\alpha_{k1} + \alpha_{k0}}$$

Likelihood Bounds & Convergence

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(x, z, v, \phi \mid \alpha, \lambda_0)] - \mathbb{E}_q[\log q(z, v, \phi)]$$

- Immediately after global parameter update, bound simplifies:

log-normalizers for cluster shape and beta stick-breaking priors

$$\mathcal{L}(q) = \mathbb{H}[r] + \sum_{k=1}^{K} [\bar{a}(\lambda_k) - \bar{a}(\lambda_0) + \log B(\alpha_{k1}, \alpha_{k0}) - \log B(1, \alpha)]$$

$$\mathbb{H}[r] = -\sum_{n=1}^{N} \sum_{k=1}^{\infty} r_{nk} \log r_{nk} = -\sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \log r_{nk}$$

For data item n = 1, 2, … N, and K candidate clusters:

$$q(z_n = k) = r_{nk} \propto e^{\mathbb{E}_q[\log \pi_k(v) + \log p(x_n | \phi_k)]}$$

For cluster k = 1, 2, … K:

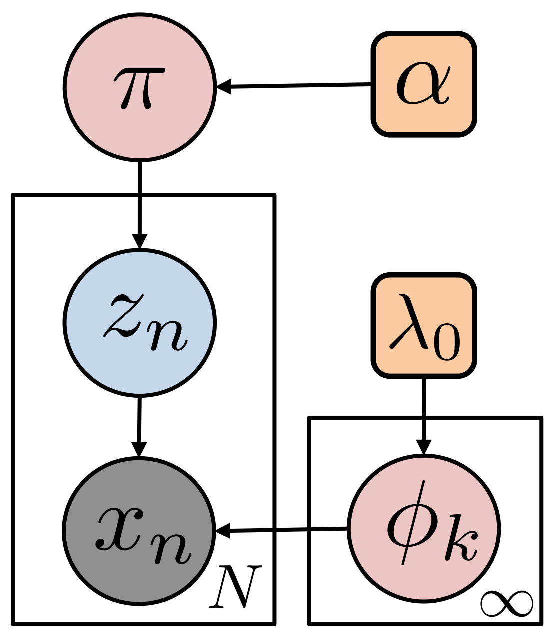
$$s_k^0 \leftarrow \sum_{n=1}^{N} r_{nk} t(x_n)$$

$$\lambda_k \leftarrow \lambda_0 + s_k^0$$

Match Expected Sufficient Statistics

$$\alpha_{k1} \leftarrow 1 + N_k^0$$

$$\alpha_{k0} \leftarrow \alpha + N_{>k}^0$$



Likelihood Bounds & Convergence

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(x, z, v, \phi \mid \alpha, \lambda_0)] - \mathbb{E}_q[\log q(z, v, \phi)]$$

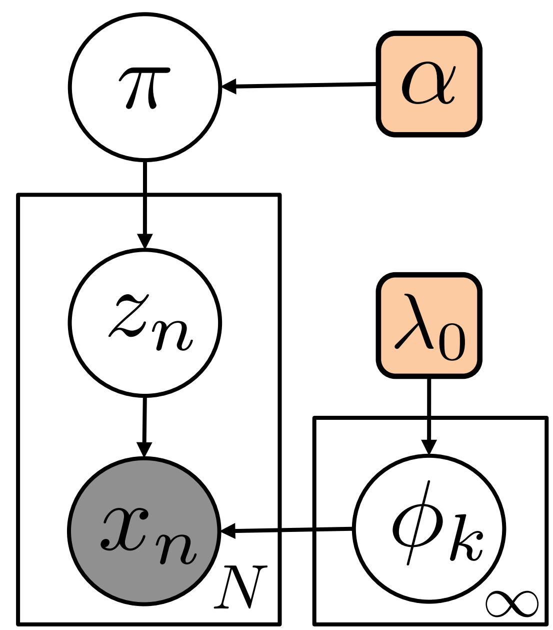
- Immediately after global parameter update, bound simplifies:

log-normalizers for cluster shape and beta stick-breaking priors

$$\mathcal{L}(q) = \mathbb{H}[r] + \sum_{k=1}^{K} [\bar{a}(\lambda_k) - \bar{a}(\lambda_0) + \log B(\alpha_{k1}, \alpha_{k0}) - \log B(1, \alpha)]$$

$$\mathbb{H}[r] = -\sum_{n=1}^{N} \sum_{k=1}^{\infty} r_{nk} \log r_{nk} = -\sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \log r_{nk}$$

- Properties of variational optimization algorithm:
  - \+ Likelihood bound monotonically increasing, guaranteed convergence to posterior mode
  - \+ Unlike classical EM for MAP estimation, allows Bayesian comparison of hypotheses with varying complexity *K*, crucial for BNP models
  - – Truncation level *K* is assumed fixed
  - – Sensitive to initialization (many modes)
  - – Each iteration must examine all data (SLOW)



Stochastic Variational Inference

Hoffman, Blei, Paisley, & Wang, JMLR 2013

Stochastically partition large dataset into *B* smaller *batches*:

Update: *For each batch b*

$r(\mathcal{B}_b) \leftarrow \mathrm{Estep}(x(\mathcal{B}_b), \alpha, \lambda)$

For cluster k = 1, 2, … K:

$s_k^b \leftarrow \sum_{n \in \mathcal{B}_b} r_{nk} t(x_n)$

$\lambda_k^b \leftarrow \lambda_0 + \frac{N}{|\mathcal{B}_b|} s_k^b$

$\lambda_k \leftarrow \rho_t \lambda_k^b + (1 - \rho_t)\lambda_k$

Apply similar updates to stick weights.

batch stats give noisy estimate of (natural) gradient

Data

$x(\mathcal{B}_1)$

$x(\mathcal{B}_2)$

$\vdots$

$x(\mathcal{B}_b)$

$\vdots$

$x(\mathcal{B}_B)$

Learning Rate

$\rho_t \triangleq (\rho_0 + t)^{-\kappa}$

Robbins-Monro convergence condition:

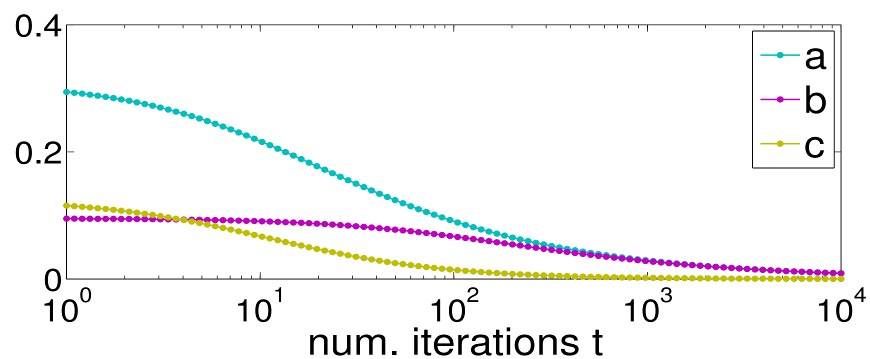
$\sum_t \rho_t \to \infty$

$\sum_t \rho_t^2 < \infty$

$\kappa \in (.5, 1]$

Properties of stochastic inference:

+ Per-iteration cost is low
+ Initial iterations often very effective



– Objective is highly non-convex, so convergence guarantee is weak
– Batch size and learning rate significantly impact efficiency & accuracy

Memoized Variational Inference

Hughes & Sudderth, NIPS 2013; Neal & Hinton 1999

Memoization: Storage (caching) of results of previous computations

Update: *For each batch b*

$r(\mathcal{B}_b) \leftarrow \mathrm{Estep}(x(\mathcal{B}_b), \alpha, \lambda)$

For cluster k = 1, 2, … K:

$s_k^0 \leftarrow s_k^0 - s_k^b$

$s_k^b \leftarrow \sum_{n \in \mathcal{B}_b} r_{nk} t(x_n)$

$s_k^0 \leftarrow s_k^0 + s_k^b$

$\lambda_k \leftarrow \lambda_0 + s_k^0$

Apply similar updates to stick weights.

batch stats allow exact estimation from partial E-steps

Data

$x(\mathcal{B}_1)$

$x(\mathcal{B}_2)$

$\vdots$

$x(\mathcal{B}_b)$

$\vdots$

$x(\mathcal{B}_B)$

Batch Summaries

$$\begin{array}{cccc} s_1^1 & s_2^1 & \cdots & s_K^1 \\ s_1^2 & s_2^2 & \cdots & s_K^2 \\ \vdots & \vdots & & \vdots \\ s_1^B & s_2^B & \cdots & s_K^B \end{array}$$

Global Summary

$$\begin{array}{cccc} s_1^0 & s_2^0 & \cdots & s_K^0 \end{array}$$

$$s_k^0 = s_k^1 + s_k^2 + \ldots s_k^B$$

Entropy for $\mathcal{L}(q)$

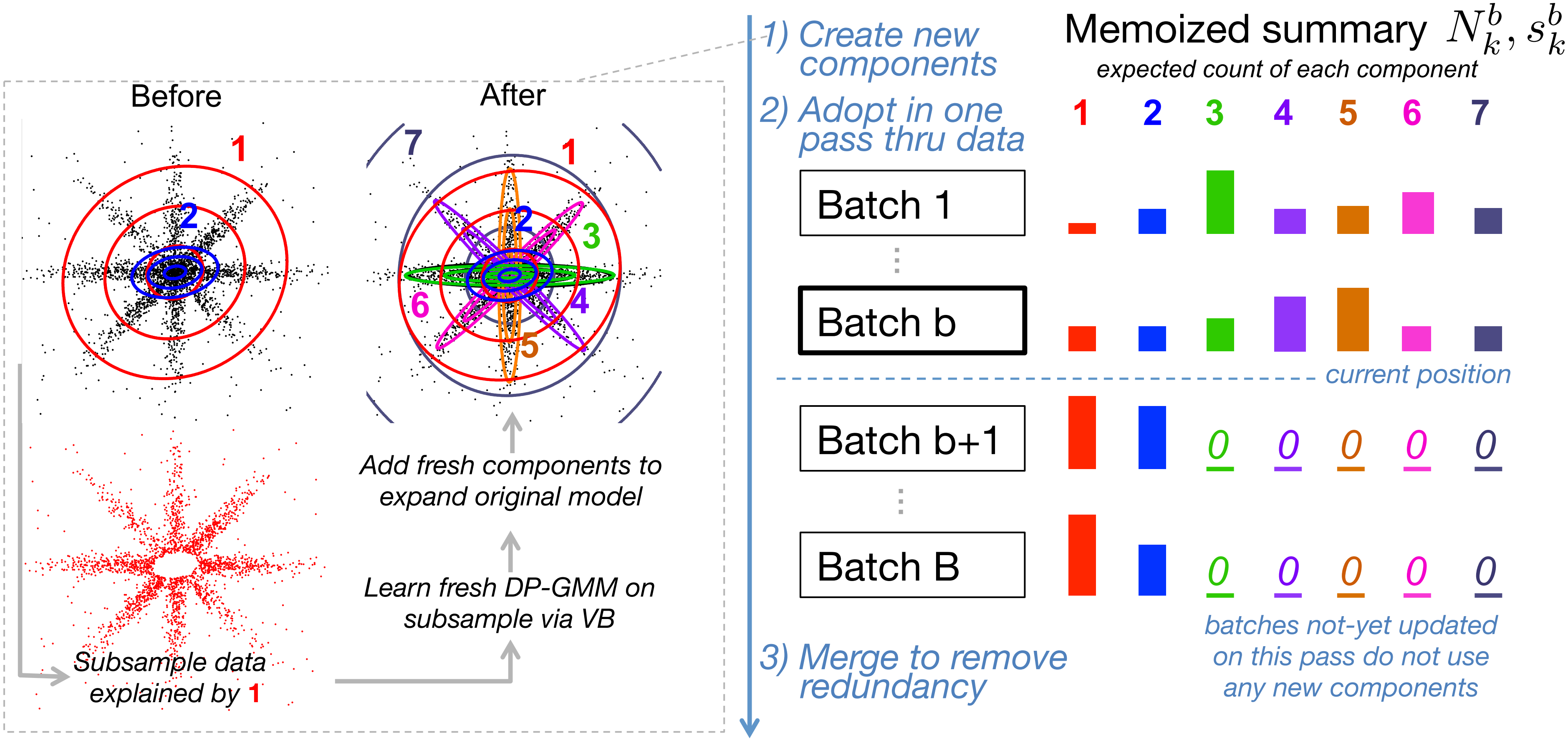
$$H_k^0 = H_k^1 + H_k^2 + \ldots H_k^B$$

$$H_k^b = -\sum_{n \in \mathcal{B}_b} r_{nk} \log r_{nk}$$

Properties of memoized inference:

- \+ Per-iteration cost is low
- \+ Initial iterations often very effective
- \+ Insensitive to chosen *B*, no learning rate
- \+ Foundation for inferring number of clusters *K*
- – Requires storage proportional to number of batches (NOT number of observations)

Memoized Cluster Births

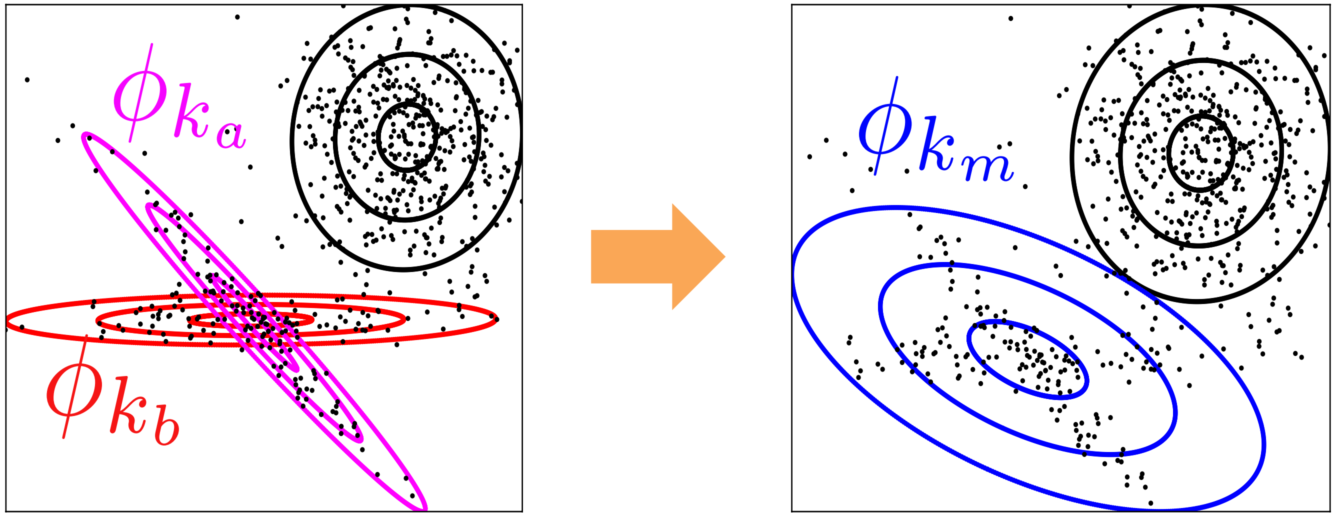


Principles guiding memoized births:

- BNP models support rare clusters, so random sampling ineffective
- Target data grouped by some current cluster (likelihood-independent)
- Memoized updates allow efficient marginal likelihood verification

Memoized Cluster Merges

Merge two clusters into one for parsimony, accuracy, efficiency.



- New cluster takes over all responsibility for data assigned to old clusters:

$$r_{nk_m} \leftarrow r_{nk_a} + r_{nk_b} \quad \Rightarrow \quad N^0_{k_m} \leftarrow N^0_{k_a} + N^0_{k_b}, \; s^0_{k_m} \leftarrow s^0_{k_a} + s^0_{k_b}$$

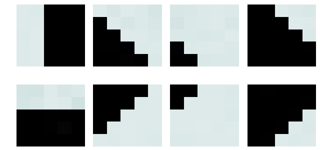
- No batch processing required, efficiently evaluate via *memoized* statistics:

$$\mathcal{L}(q) = \mathbb{H}[r] + \sum_{k=1}^{K} \left[\bar{a}(s^0_k + \lambda_0) - \bar{a}(\lambda_0) + \log B(1 + N^0_k, \alpha + N^0_{>k}) - \log B(1, \alpha) \right]$$

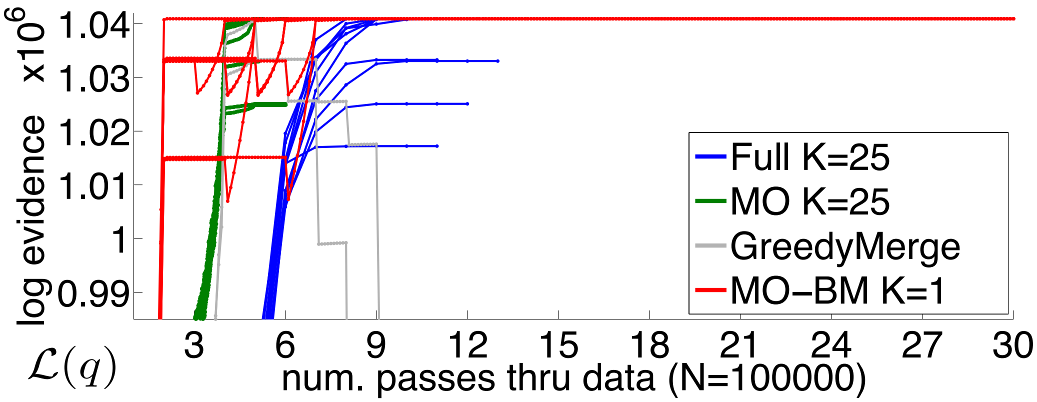
- Accept or reject via *exact* full-dataset likelihood bound: $\mathcal{L}(q_{\text{merge}}) > \mathcal{L}(q)$?

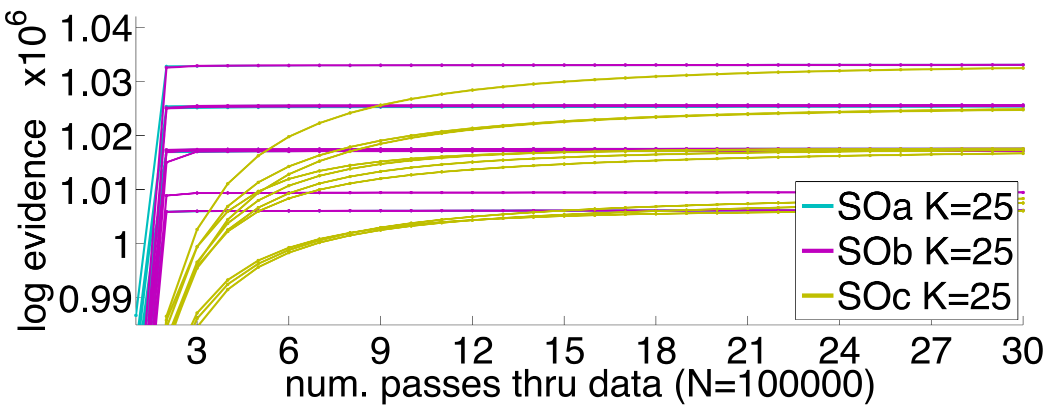
  Requires memoized entropy sums for candidate pairs of clusters;
  more efficient alternatives under development.

Example: Finite Gaussian Mixture



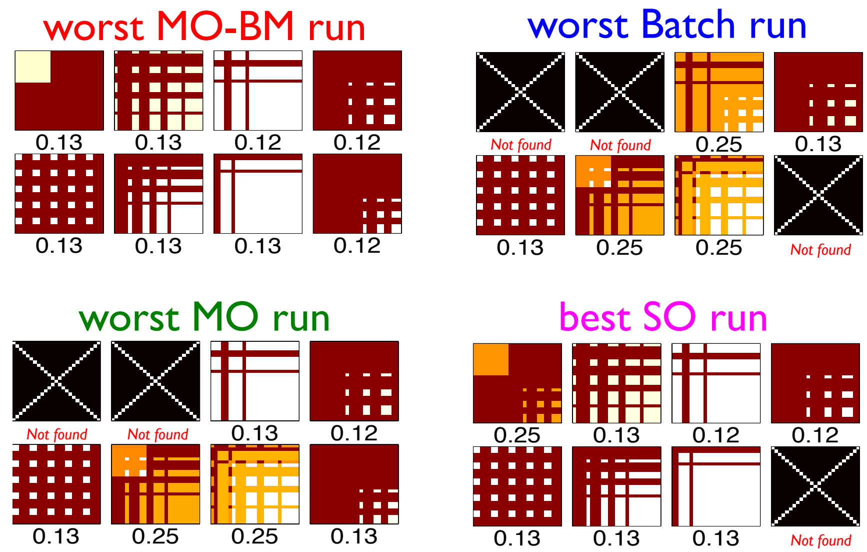
- *N=100,000* samples from mixture of *8* Gaussians
- *25-dim.* covariance motivated by *5x5* image patches
- DP mixture variational approximations allow *K=25* clusters





Batch, memoized, & memoized birth-merge
Stochastic variational: Rate a, Rate b, Rate c
Greedy: Merge based on single batches

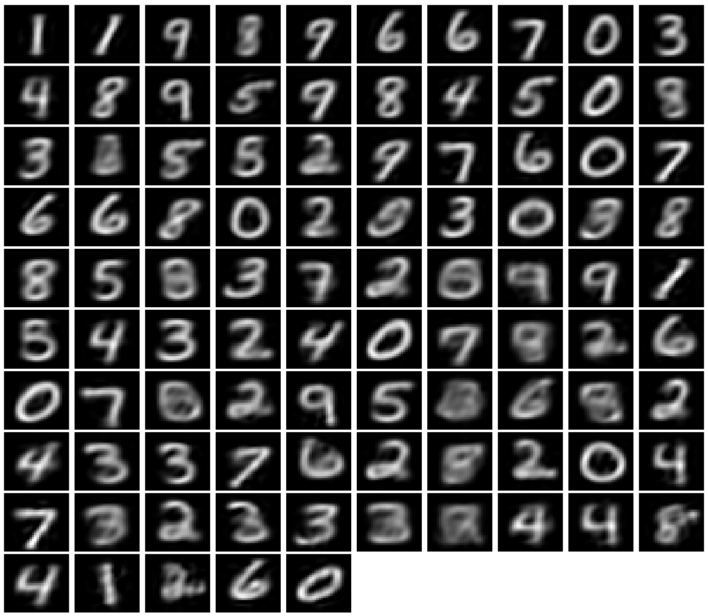
From 10 random initializations:



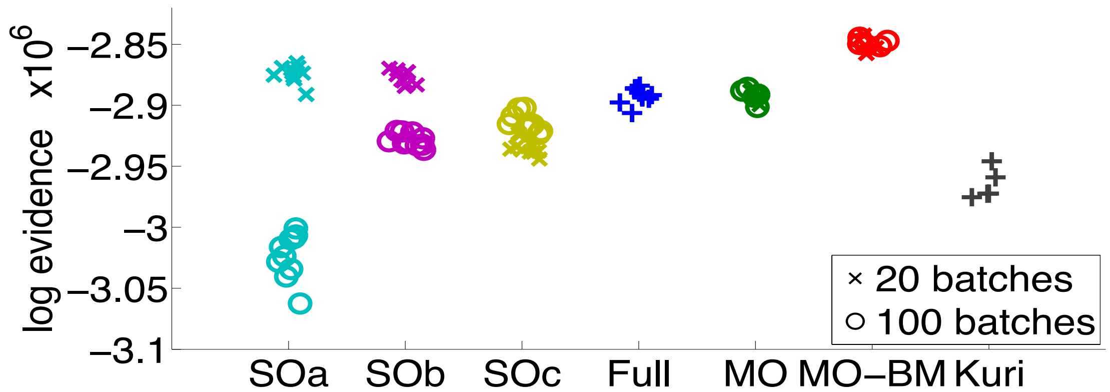
- Memoized birth-merge from K=1 finds true cluster every time
- Greedy decisions to merge based on single batches collapse model
- Stochastic sensitive to learning rate and initialization

Clustering Handwritten Digits

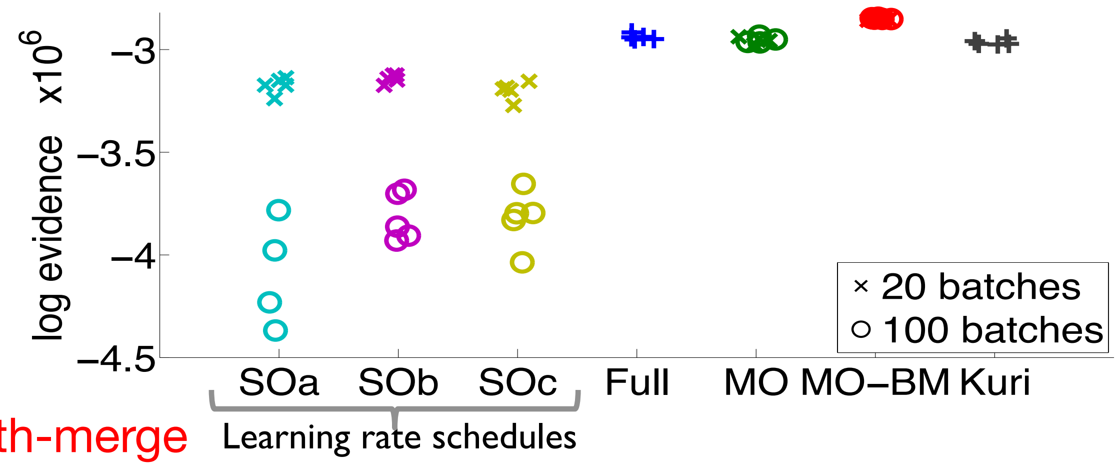
MNIST: *60,000 digits projected to 50 dimensions via PCA.*



Likelihood bound, K-means++ initialization



Likelihood Bound, random initialization



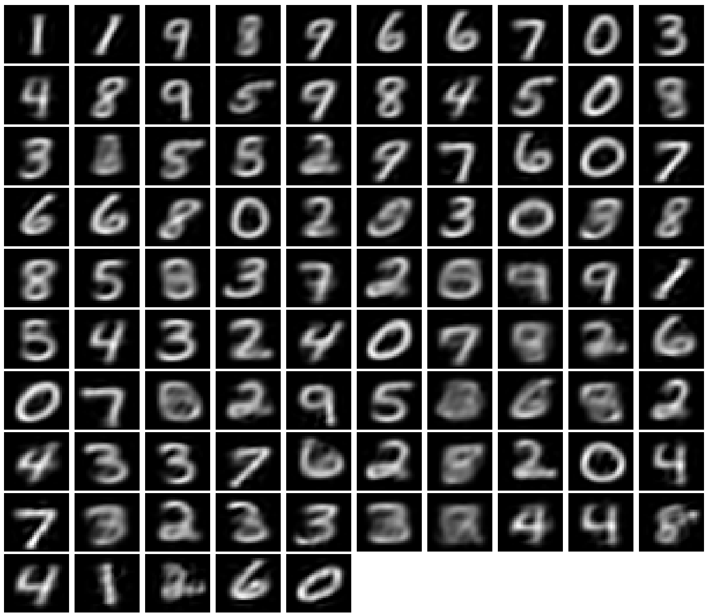
Batch, memoized, & memoized birth-merge

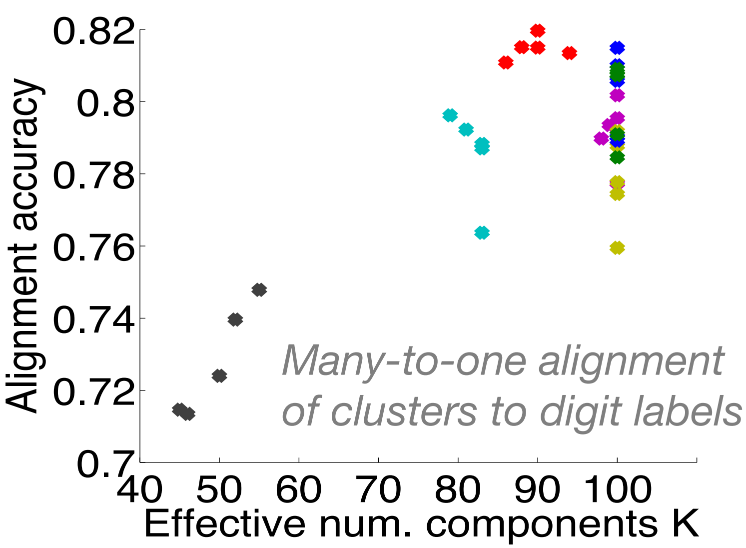
Stochastic variational: Rate a, Rate b, Rate c

Kurihara: Accelerated variational, NIPS 2006

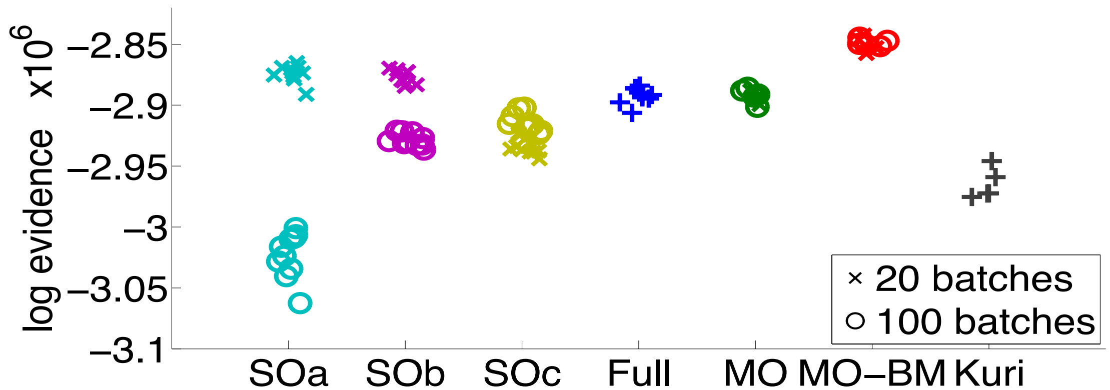
Clustering Handwritten Digits

MNIST: *60,000 digits projected to 50 dimensions via PCA.*

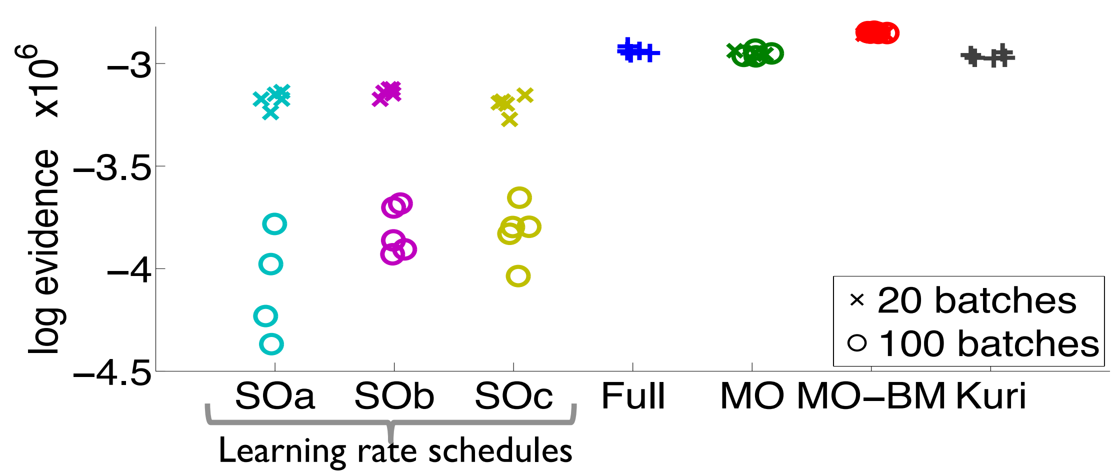




Likelihood bound, K-means++ initialization

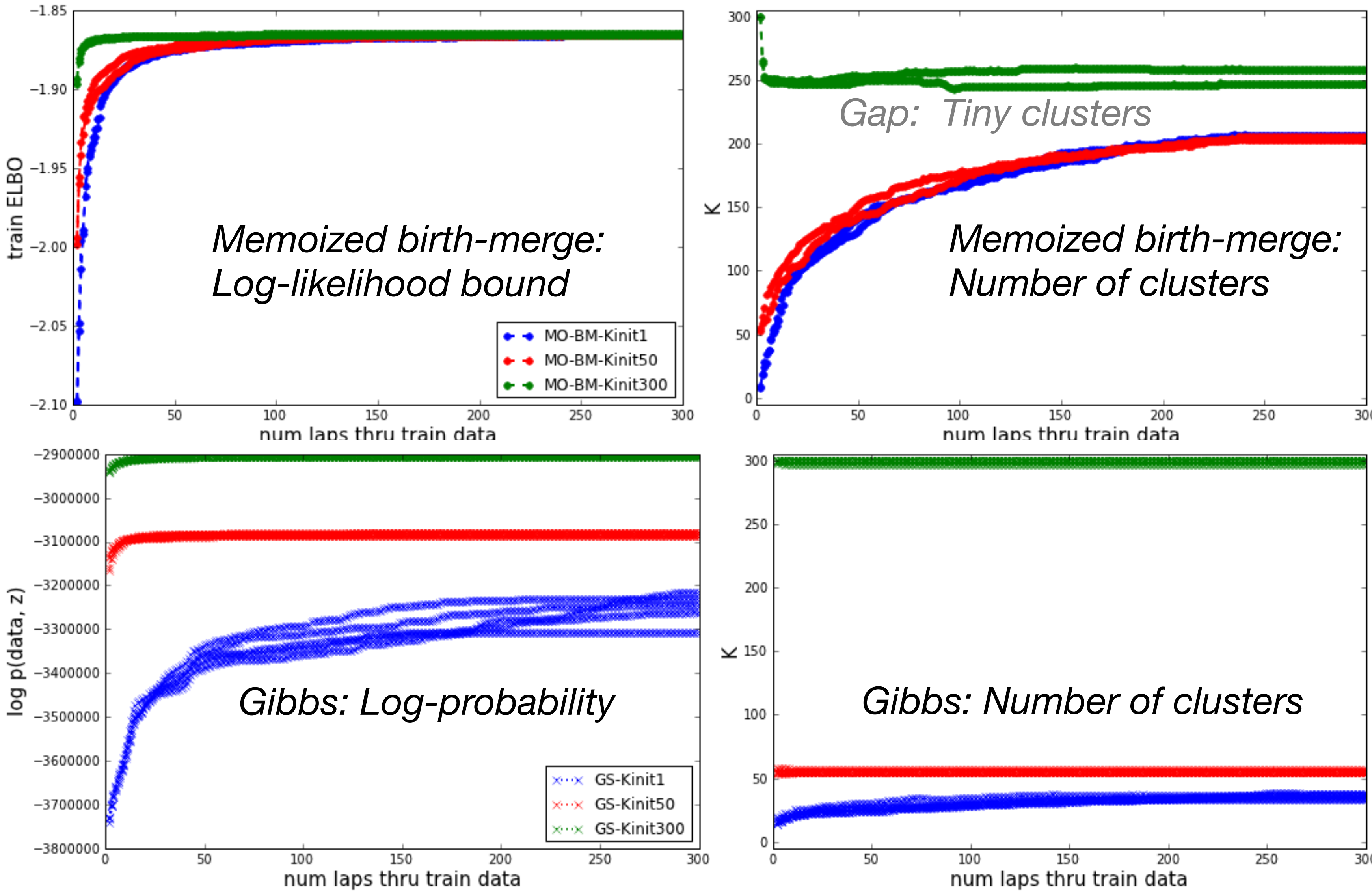


Likelihood Bound, random initialization



- Memoized birth-merge from K=1 has highest accuracy while using fewer clusters

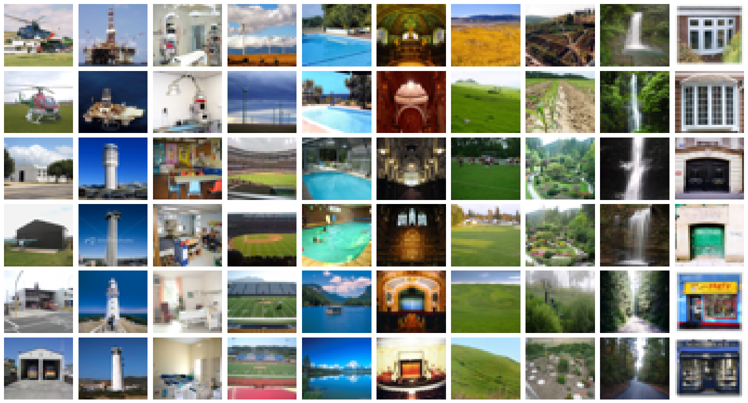
MNIST: Variational versus Gibbs

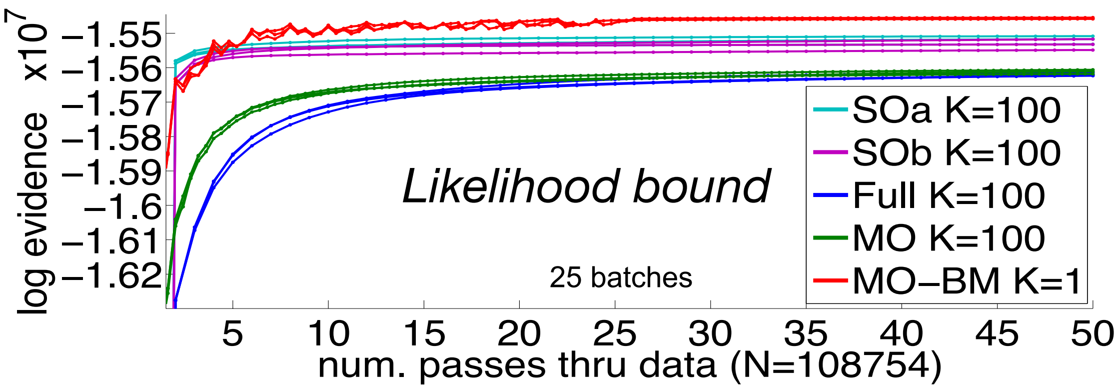


- Five random initializations from *K=1*, *K=50*, *K=300* clusters
- Diagonal-covariance Gaussians (change from previous slides)

Clustering Image Patches

SUN Database of Natural Scene Categories: *N=108,754*

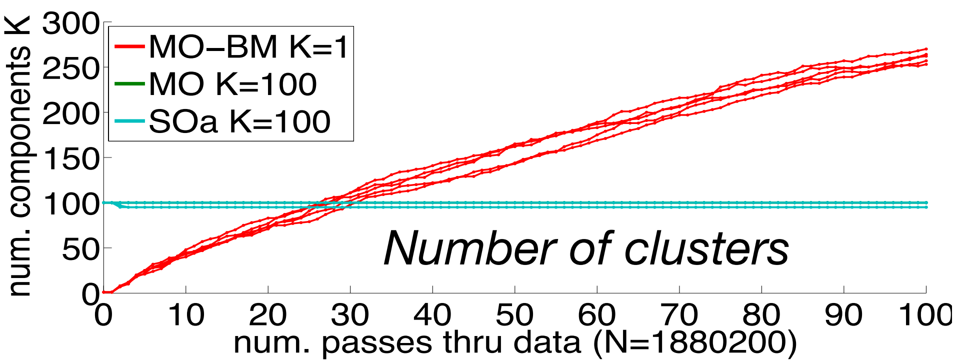


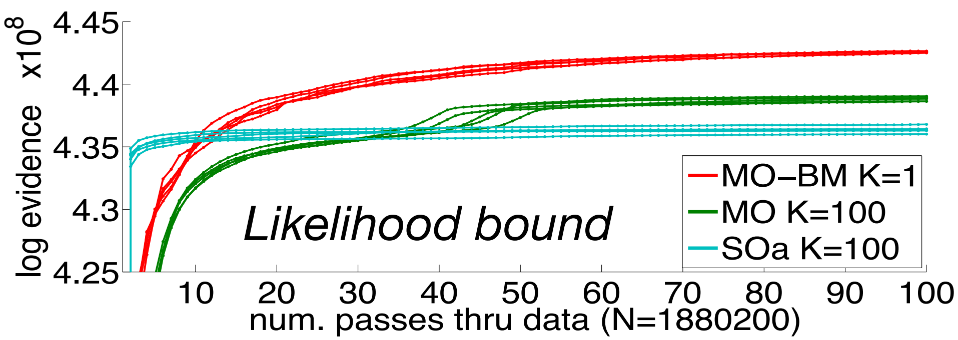


- Memoized birth-merge learns a more accurate model with only *K=28* clusters

8x8 Image Patches (Berkeley Segmentation): *N=1.88 million*

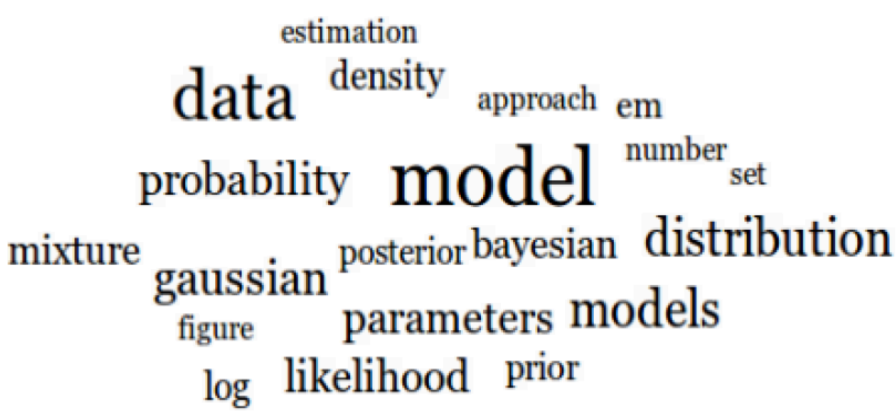
- Memoized birth-merge allows growth in model complexity
- Effective performance as density model for image denoising





Memoized Variational Inference for Hierarchial DP Topic Models

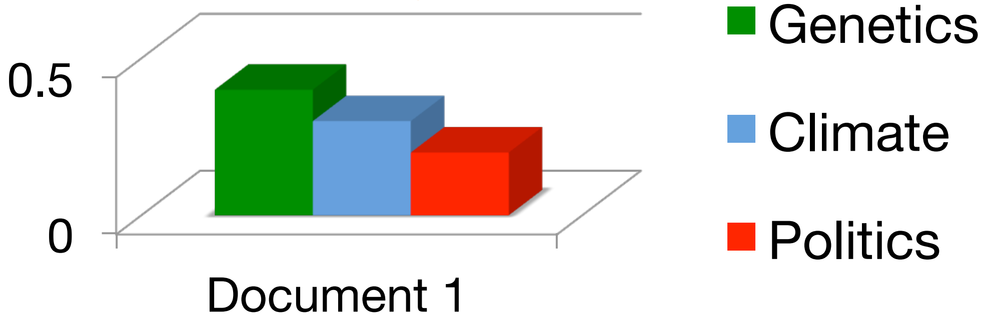
Michael Hughes, Dae Il Kim, & E. Sudderth



What are Topic Models?

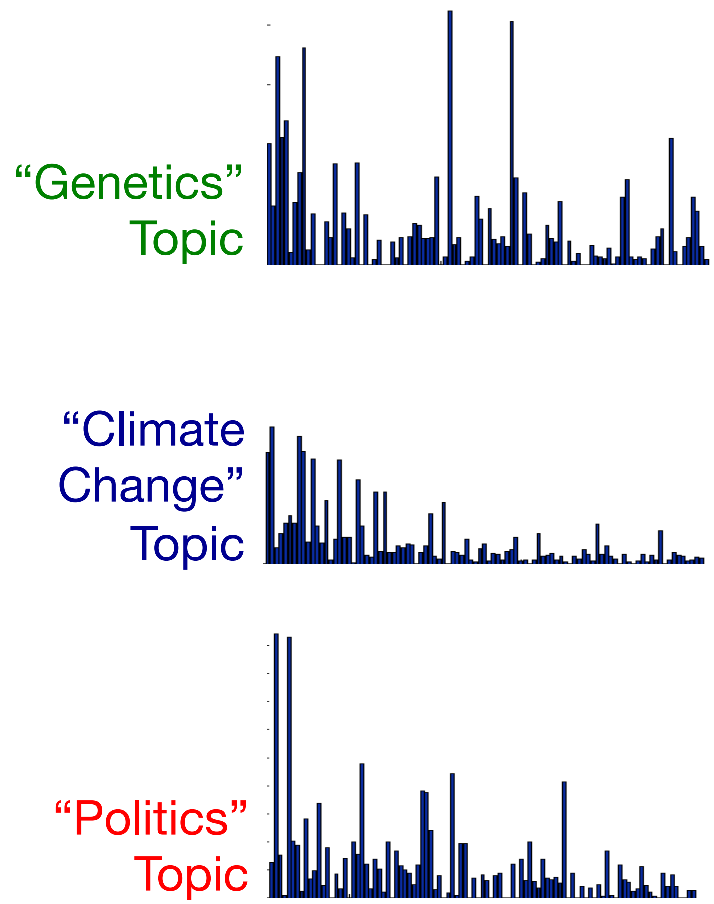
GOAL: Summarize semantic content of a large document corpus.

There are reasons to believe that the genetics of an organism are likely to shift due to the extreme changes in our climate. To protect them, our politicians must pass environmental legislation that can protect our future species from becoming extinct…



Documents are represented as mixtures of "topics" used with varying frequencies.

Topics are categorical distributions on a (typically large) discrete vocabulary:



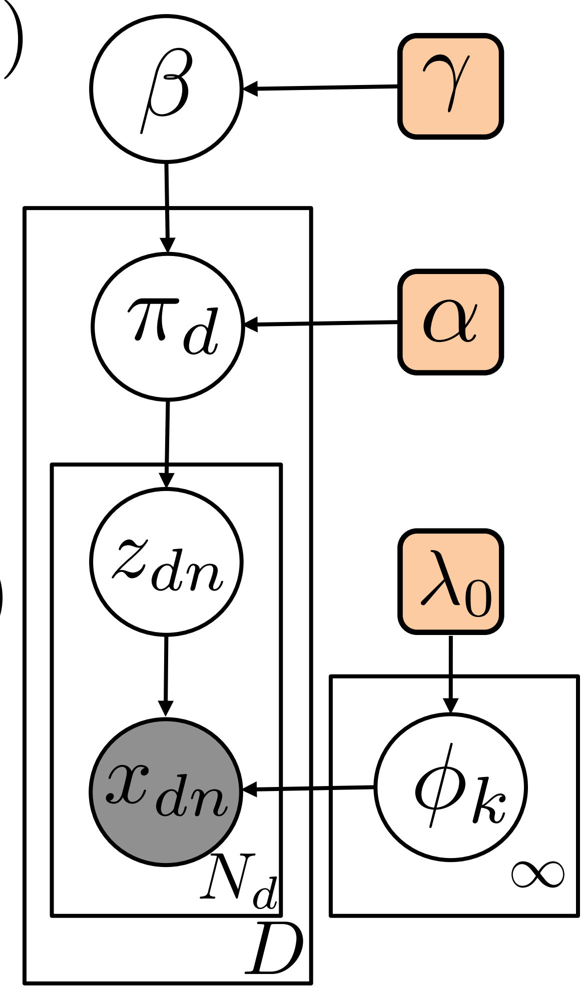
Hierarchical DP Topic Model

Generalization of Latent Dirichlet Allocation (LDA, Blei 2003) by Teh et al. JMLR 2006.
Dependent Dirichlet process (DDP, MacEachern 1999) with group-specific weights.

- Global topic frequencies and parameters:

$$\beta_k = u_k \prod_{\ell=1}^{k-1}(1-u_\ell) \qquad u_k \sim \mathrm{Beta}(1,\gamma)$$

$$\phi_k \sim \mathrm{Dirichlet}(\lambda_0)$$ *(sparse)*



- For each of *D* documents (groups):
  - Topic frequencies: $\pi_d \sim \mathrm{DP}(\alpha\beta)$

Generalized Dirichlet, Connor & Mosimann 1969

$$\pi_{dk} = v_{dk} \prod_{\ell=1}^{k-1}(1-v_{d\ell})$$

$$v_{dk} \sim \mathrm{Beta}(\alpha_k u_k, \alpha_k(1-u_k))$$

$$\alpha_k = \alpha \prod_{\ell=1}^{k-1}(1-v_{d\ell})$$

- For each of N_d words in document *d*:
  - Topic assignment: $z_{dn} \sim \mathrm{Cat}(\pi_d)$
  - Observed value: $x_{dn} \sim \mathrm{Cat}(\phi_{z_{dn}})$

Variational Learning of HDP Topics

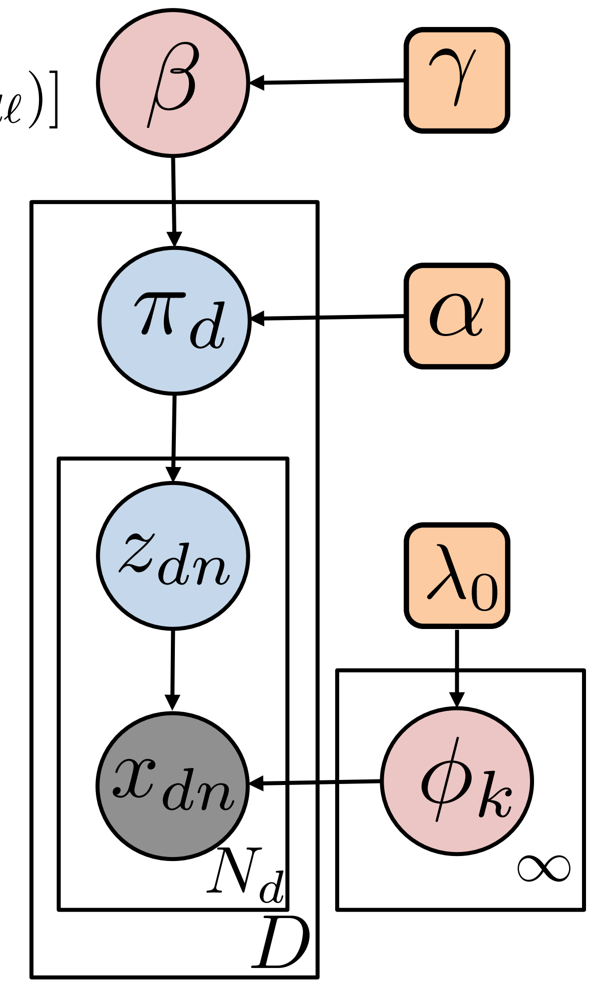
$$q(z_{dn}) = \mathrm{Cat}(z_{dn} \mid r_{dn1}, r_{dn2}, \ldots, r_{dnK}, 0, 0, 0, \ldots)$$ *for some K>0*

Update Document Distributions: For $k \leq K$,

$$r_{dnk} \propto \exp(\mathbb{E}_q[\log \pi_{dk}(v_d)] + \mathbb{E}_q[\log p(x_{dn} \mid \phi_k)])$$

$$\mathbb{E}_q[\log \pi_{dk}(v_d)] = \mathbb{E}_q[\log(v_{dk})] + \sum_{\ell=1}^{k-1} \mathbb{E}_q[\log(1 - v_{d\ell})]$$

- Closed form update for beta stick-breaking weights
- Local iteration between assignments and weights



Update Global Parameters:

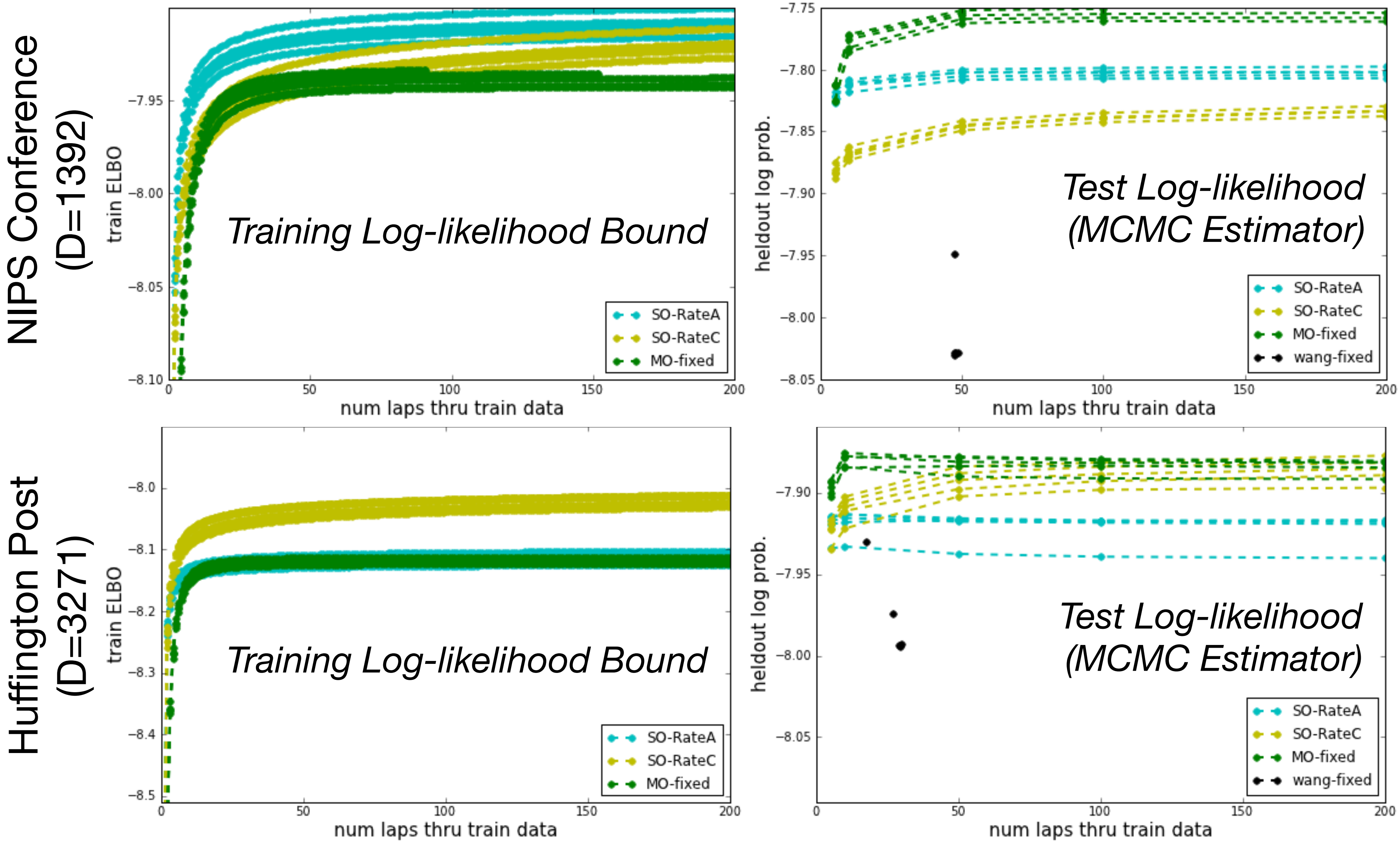
$$q(\phi_k) = \mathrm{Dir}(\phi_k \mid \lambda_0 + s_k^0)$$

$$s_k^0 \leftarrow \sum_{d=1}^{D} \sum_{n=1}^{N_d} r_{dnk} t(x_{dn})$$

- Closed form for topic-specific word distributions
- Beta normalization constants have non-conjugate dependence on topic frequencies, requires additional bound and numerical optimization

Iterate: Batch, Stochastic, or Memoized

Analysis of Document Corpora

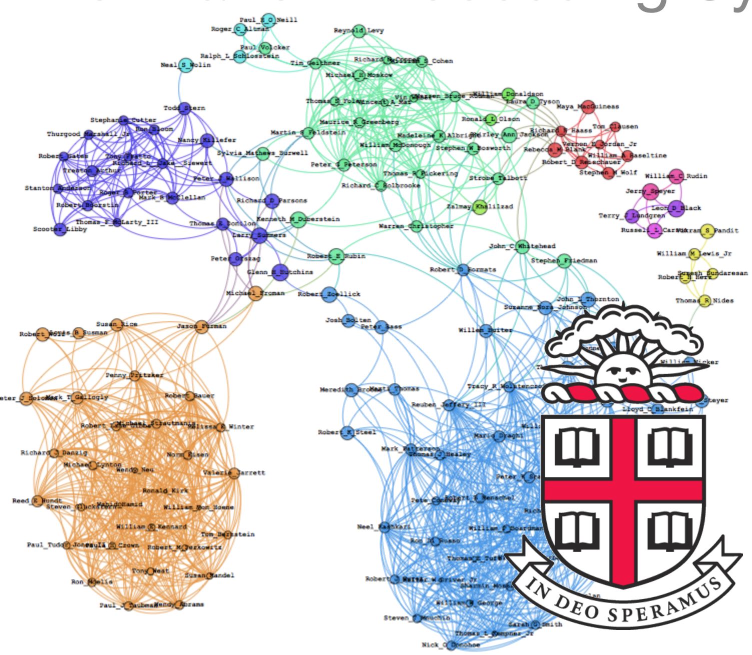


- ➢ Variational: Memoized versus stochastic rate A, stochastic rate C
- ➢ Baseline: Stochastic variational on “expanded” HDP (Wang et al. 2011)

Stochastic Variational Inference for Hierarchial DP Relational Models

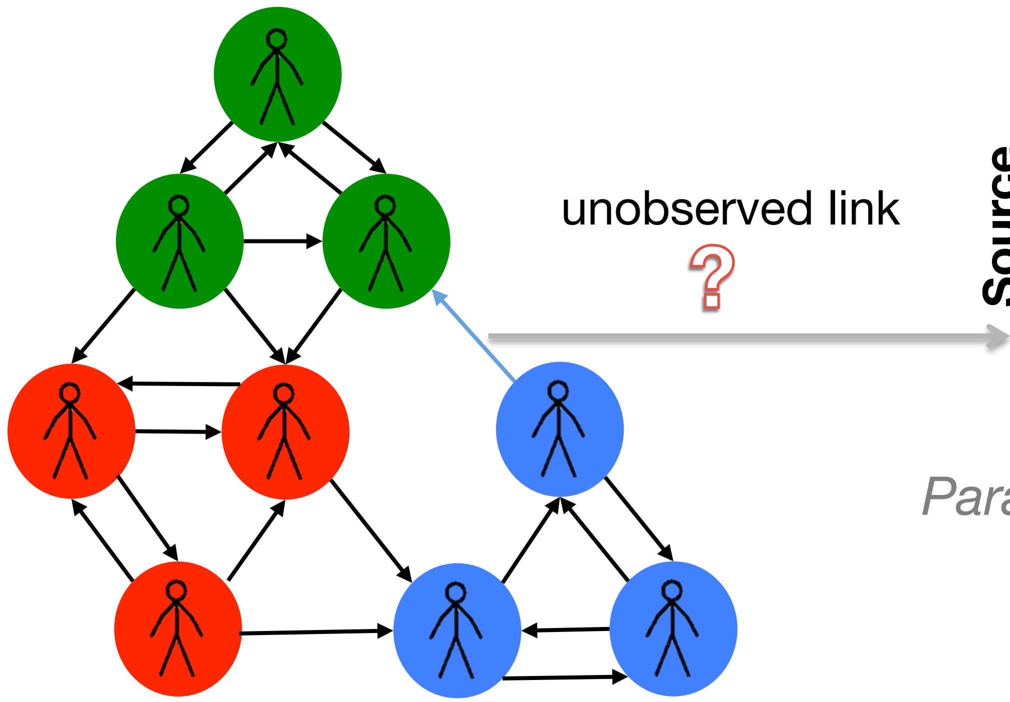
D. Kim, P. Gopalan, D. Blei, & E. Sudderth

2013 Conference on Neural Information Processing Systems



What are Relational Models?

GOAL: *Unsupervised community discovery from observed relationships.*



	Receiver		
Source			
	.8	.6	.2
	.1	.9	.3
	.1	.2	.8

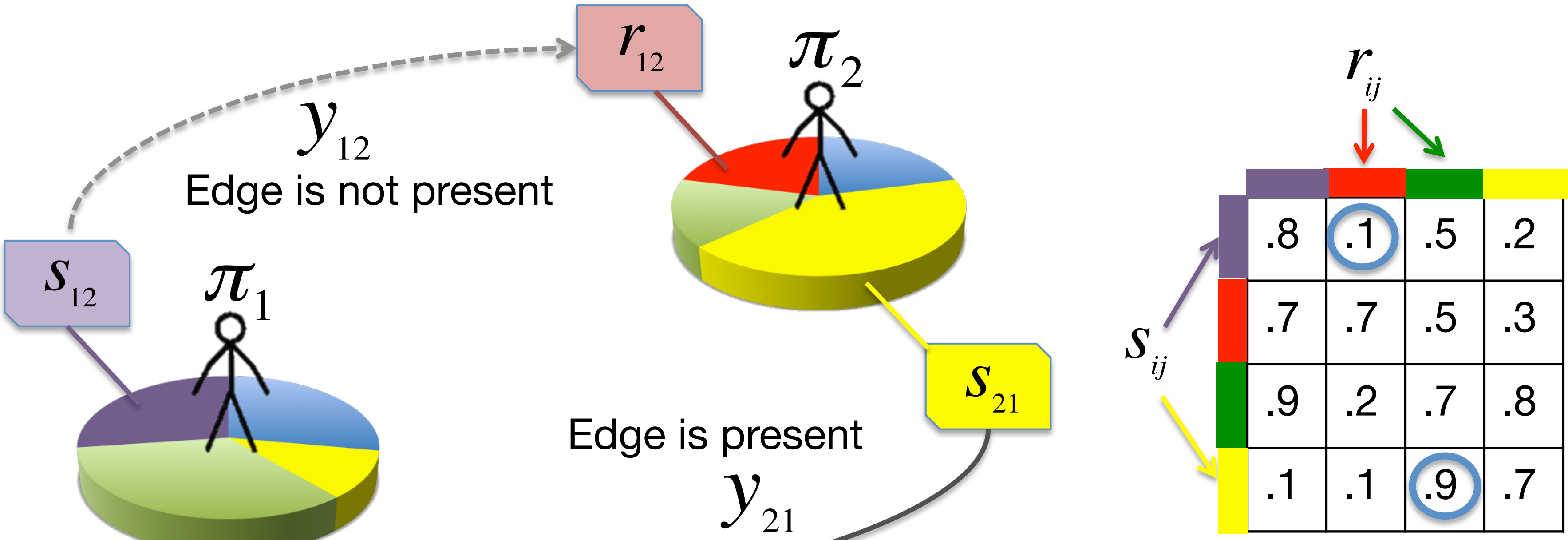
Edge Creation Parameter Matrix (K=3)

Stochastic Block Model: *(Wang et al., JASA 1987)*

- Assign each node to *one* latent block/community
- Predict edge presence or absence from block assignments of *source* and *receiver* nodes

Mixed Membership Blockmodels

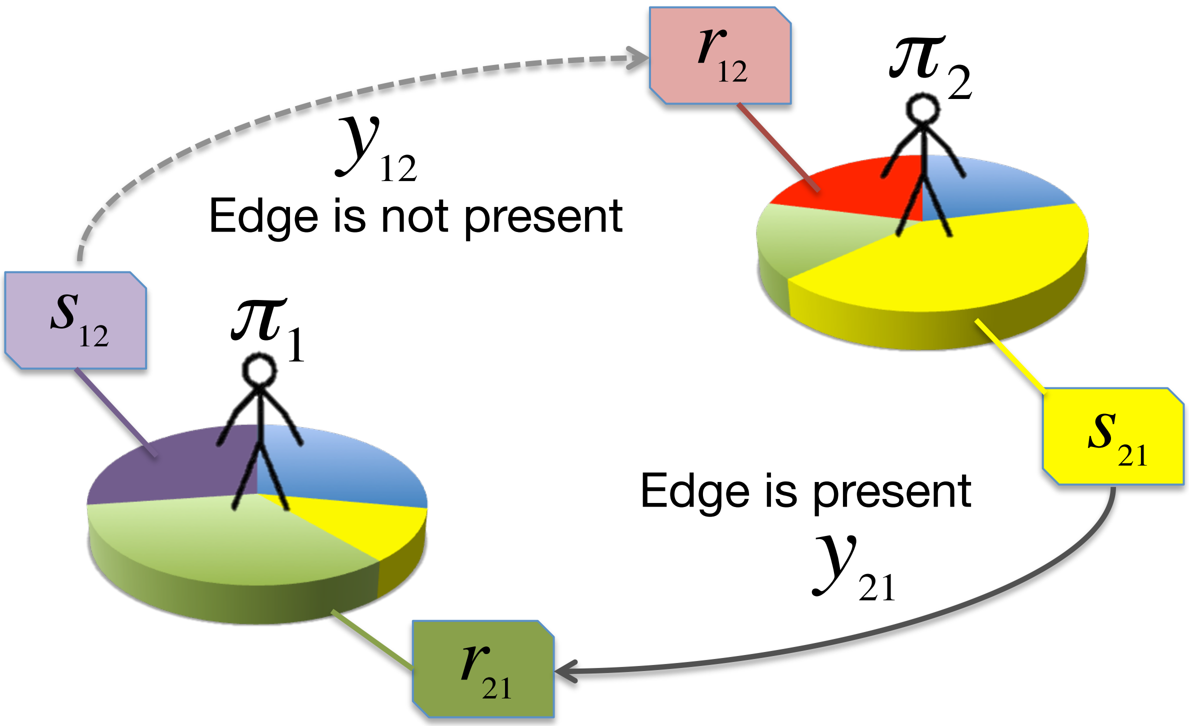
Parametric mixed membership stochastic blockmodel, Airoldi et al. JMLR 2008



Edge Creation Parameter Matrix

- Source Community Assignment — $s_{ij} \sim \mathrm{Cat}(\pi_i)$
- Receiver Community Assignment — $r_{ij} \sim \mathrm{Cat}(\pi_j)$
- Community Link Probability — $\phi_{k\ell} \sim \mathrm{Beta}(\tau_a, \tau_b)$
- Binary Edge Indicator — $y_{ij} \sim \mathrm{Bern}(s_{ij} \phi r_{ij}^T)$

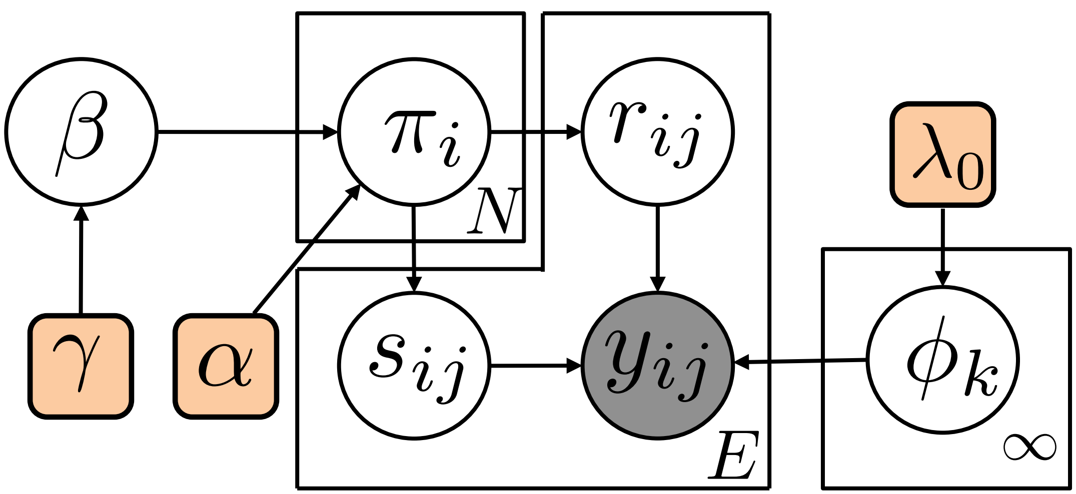
HDP Relational Models



r_{ij} (columns), s_{ij} (rows)

.8	.1	.5	.2
.7	.7	.5	.3
.9	.2	.7	.8
.1	.1	.9	.7

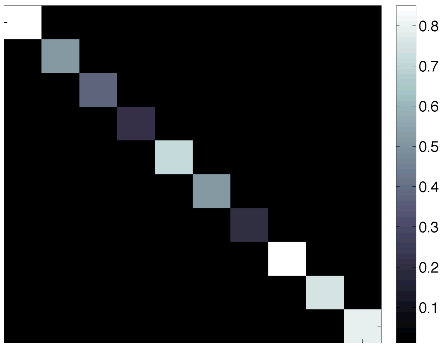
Edge Creation Parameter Matrix



$$
\begin{aligned}
s_{ij} &\sim \mathrm{Cat}(\pi_i) \\
r_{ij} &\sim \mathrm{Cat}(\pi_j) \\
\phi_{k\ell} &\sim \mathrm{Beta}(\tau_a, \tau_b) \\
y_{ij} &\sim \mathrm{Bern}(s_{ij} \phi r_{ij}^T)
\end{aligned}
$$

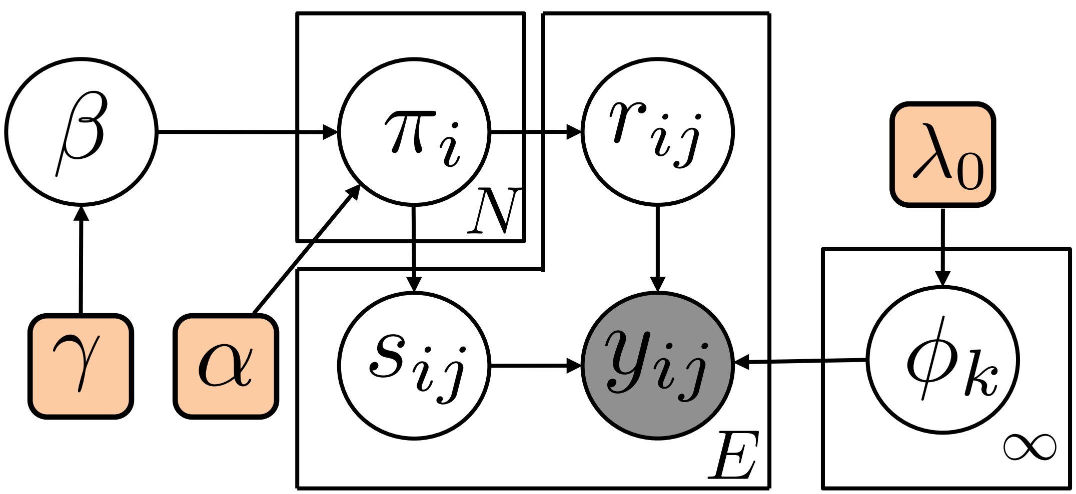
Variational Learning of Relations

Assortative Likelihoods:

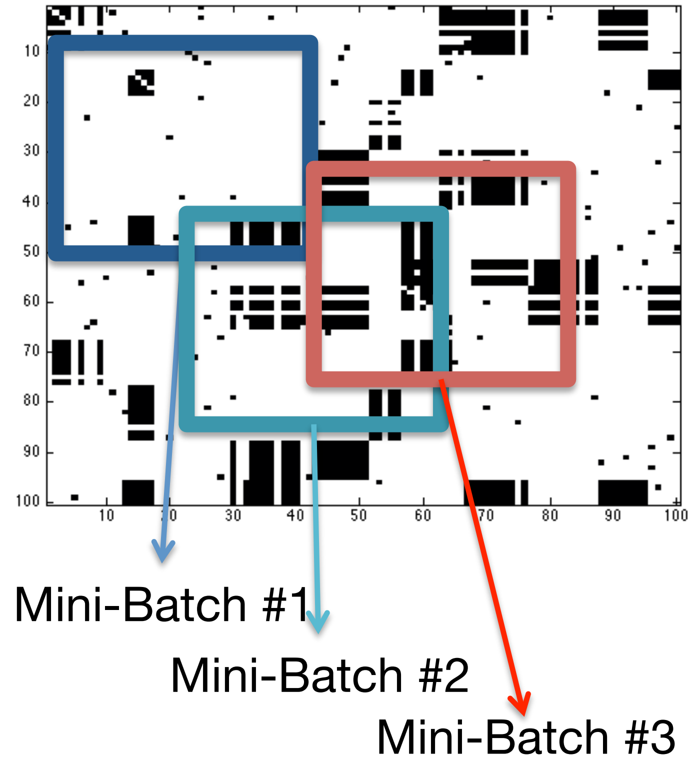


$$p(y_{ij} = 1 \mid s_{ij} = r_{ij} = k) = \phi_k$$
$$p(y_{ij} = 1 \mid s_{ij} \neq r_{ij}) = \epsilon$$

$\mathcal{O}(K)$ *storage & computation for distribution on K^2 community pairs*

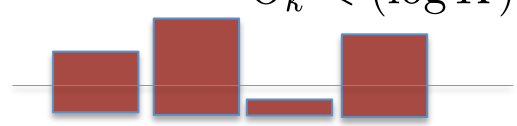


Stochastic Variational:



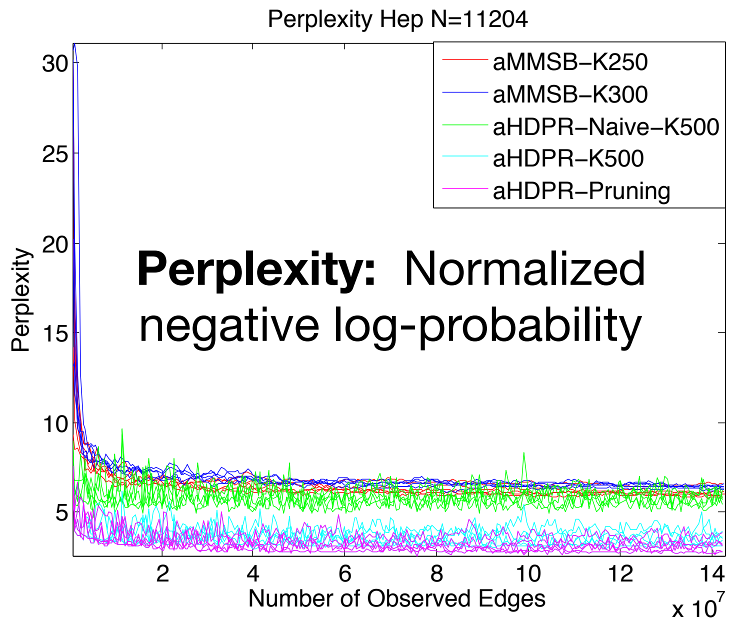
Variational Pruning:

$\Theta_k < (\log K)/N$

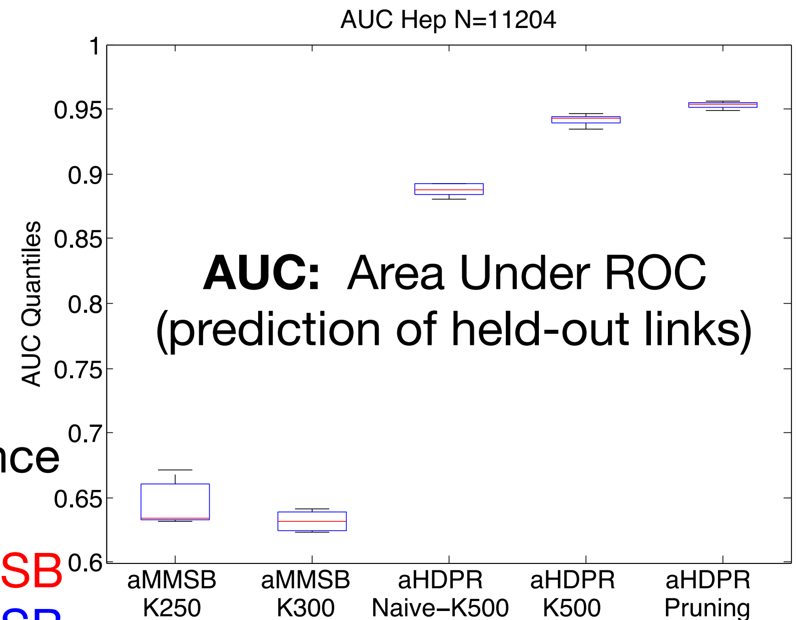


$$\Theta_k = \sum_{i=1}^{N} \mathbb{E}_q[\pi_{ik}]$$

Analysis of Collaboration Networks

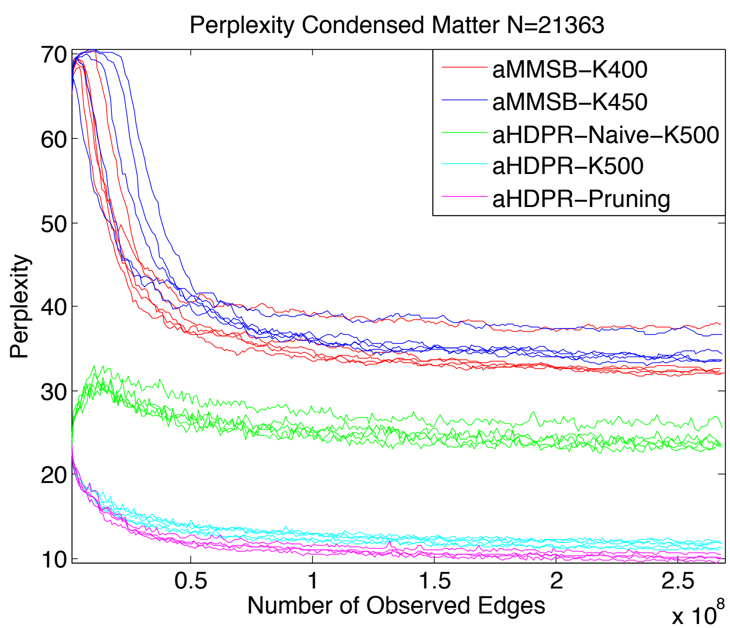


High Energy Physics (HEP) N=11,204

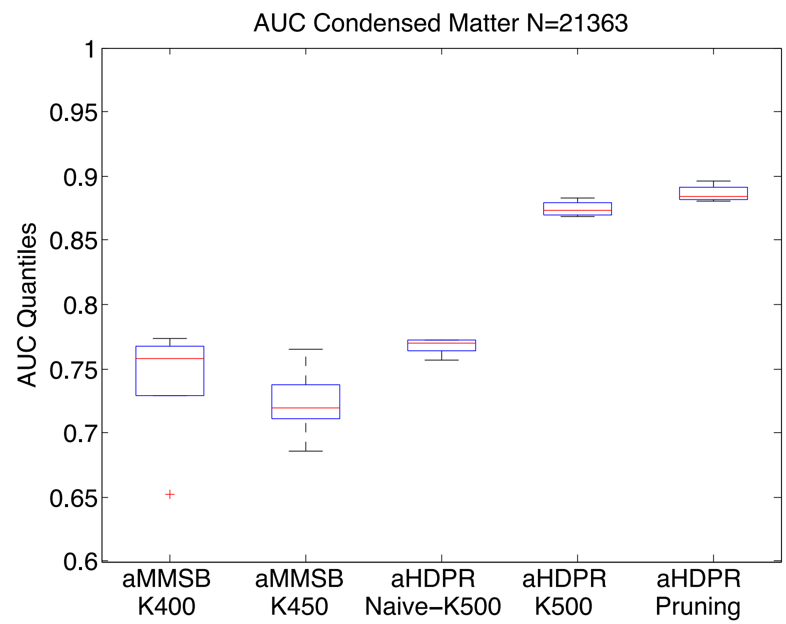


Stochastic inference & model variants:

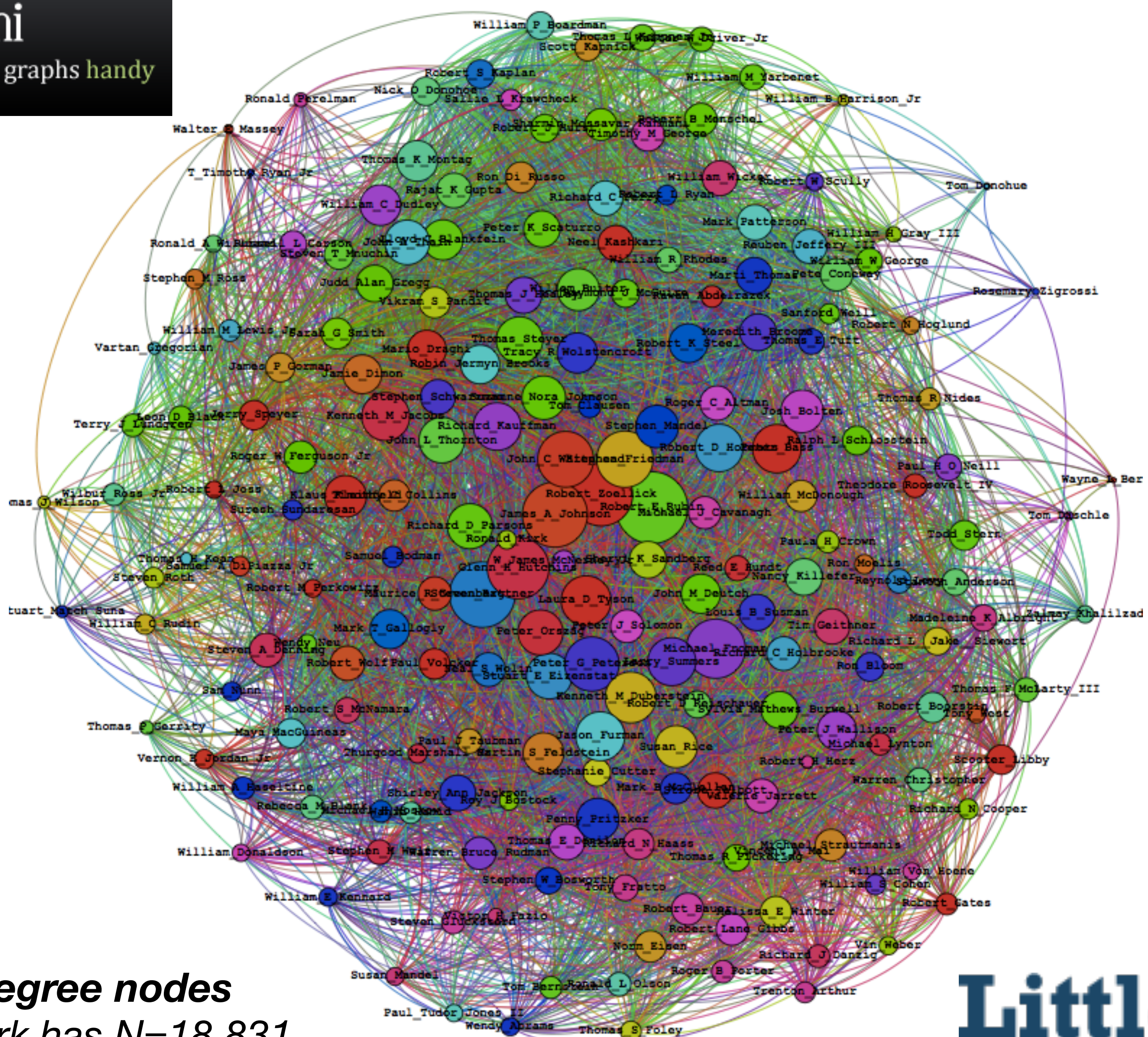
- Parametric MMSB
- Parametric MMSB
- HDP naïve
- HDP blocked
- HDP pruned



Condensed Matter Physics N=21,363



LittleSis Network: Raw Data



Top 200 degree nodes

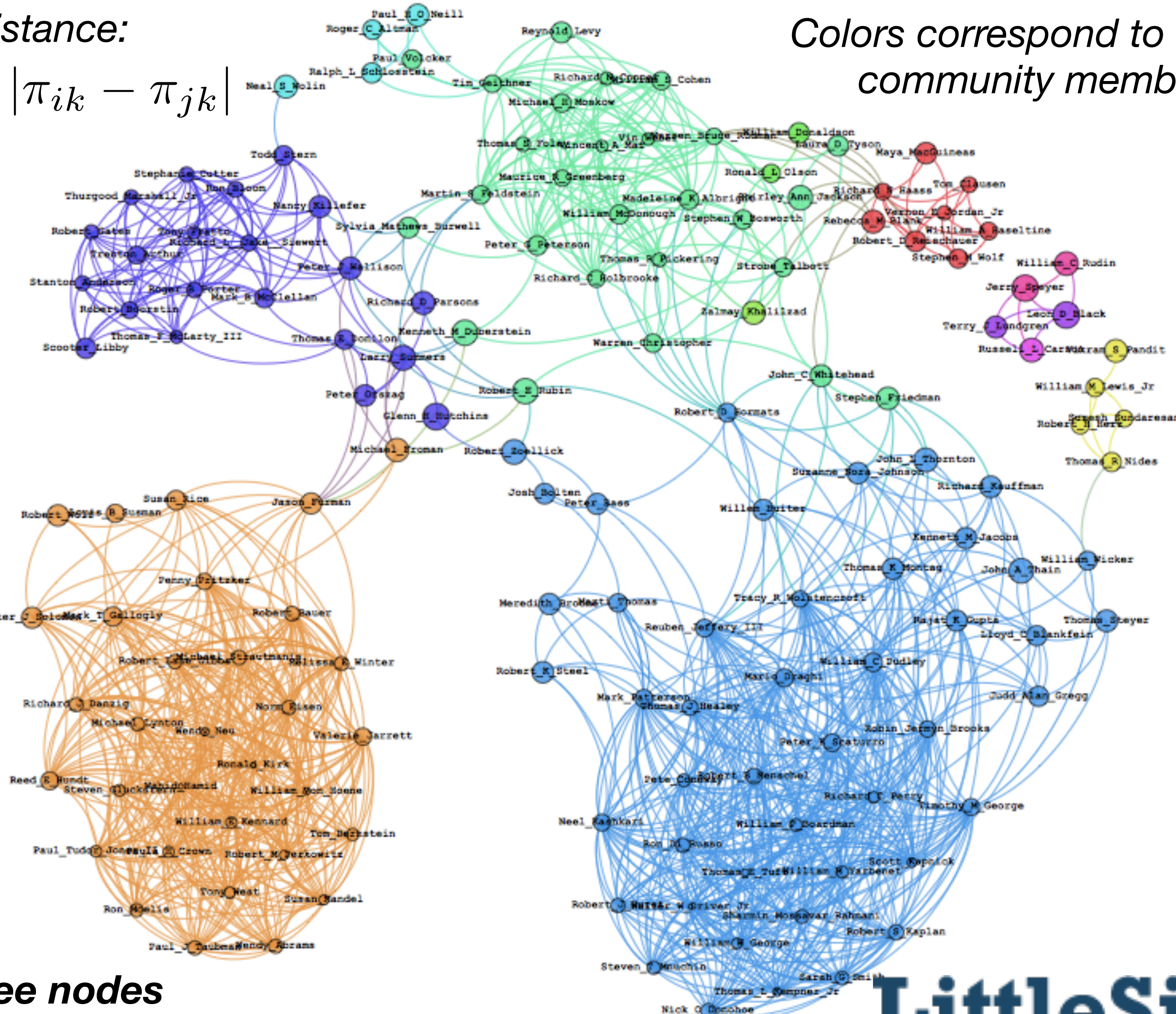
Full network has N=18,831

LittleSis Network Communities

Community Distance:

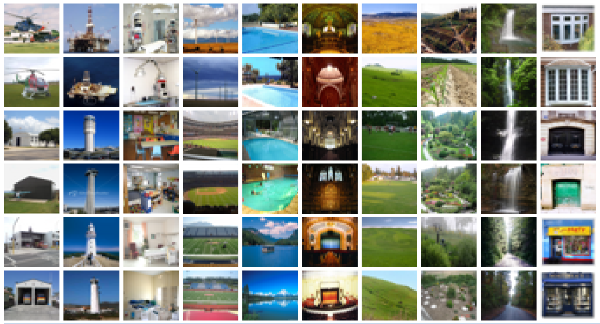
$$D_{ij} = \frac{1}{2}\sum_k |\pi_{ik} - \pi_{jk}|$$

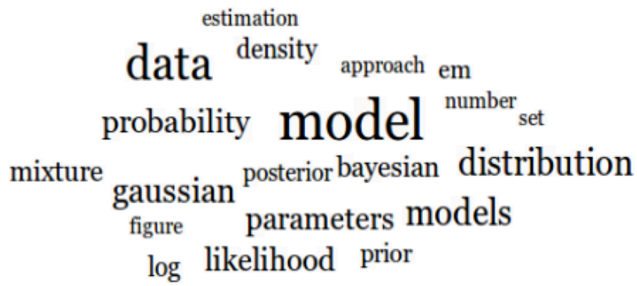
Colors correspond to highest community memberships

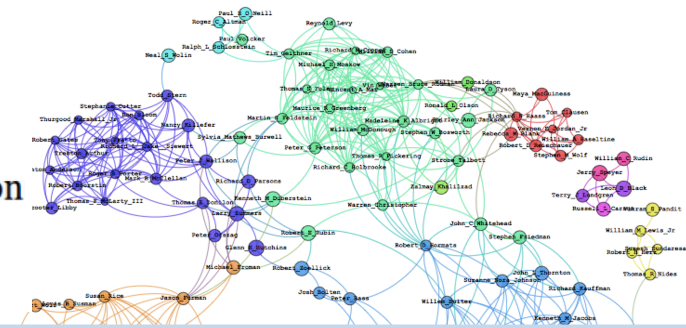


Top 200 degree nodes

Full network has N=18,831

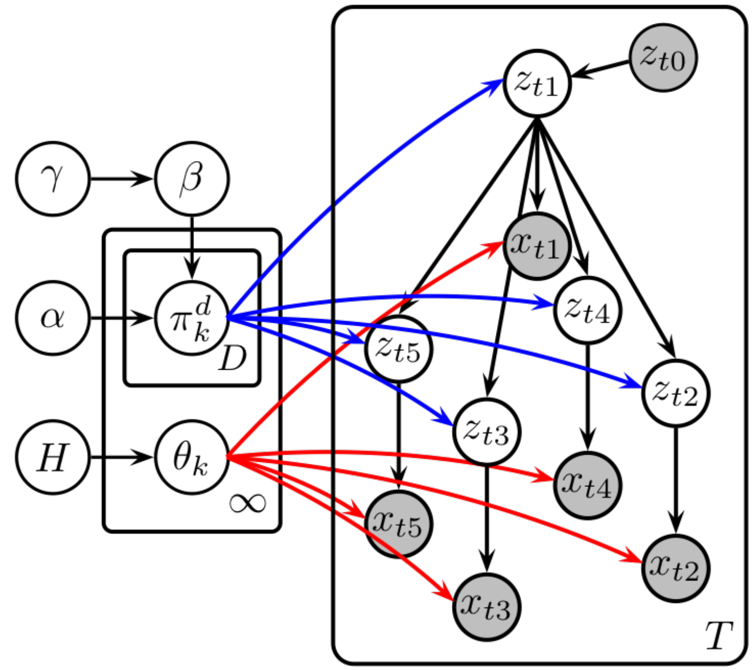






Reliable Variational Learning for Hierarchical Dirichlet Processes

- **Scalable:** Large-scale learning via stochastic or memoized updates
- **Reliable:** Birth-merge recovers structure informed by model & data, not inference algorithm limitations
- **Flexible:** Designed to be broadly applicable: space, time, scale, …



BNPy: Bayesian Nonparametric Learning in Python

Erik Sudderth @ Brown CS: http://cs.brown.edu/~sudderth/