Nonparametric Bayesian Models
--Learning and Reasoning in Open Possible Worlds

Eric Xing
epxing@cs.cmu.edu
Machine Learning Dept./Language Technology Inst./Computer Science Dept.
Carnegie Mellon University

Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Transformed DP and HDP
  - Kernel stick-breaking process
  - Application: joint image segmentation
- Dynamic Dirichlet Process
  - Hidden Markov DP
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary

Clustering

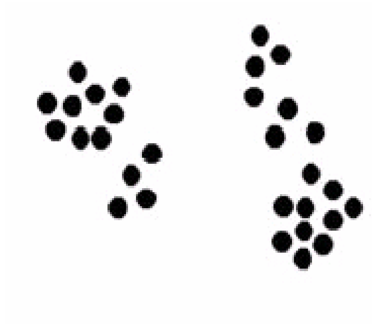
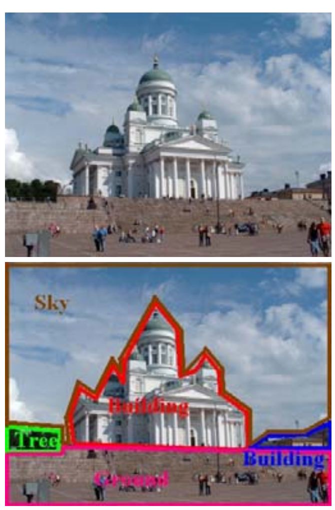


Image Segmentation



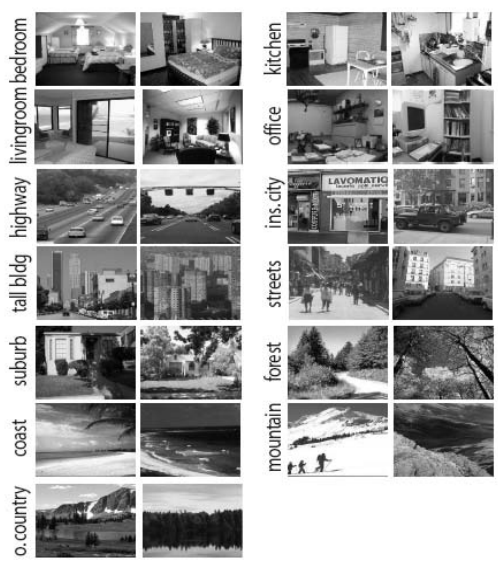
- How to segment images?
  - Manual segmentation (very expensive)
  - Algorithm segmentation
    - K-means
    - Statistical mixture models
    - Spectral clustering
- Problems with most existing algorithms
  - Ignore the spatial information
  - Perform the segmentation one image at a time
  - Need to specify the number of segments *a priori*

Discover Object Categories

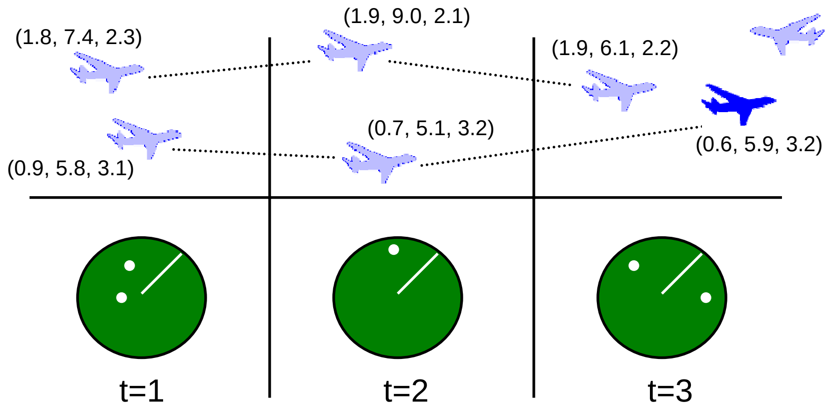
- Discover what objects are present in a collection of images in an unsupervised way

- Find those same objects in novel images
- Determine what local image features correspond to what objects; segmenting the image

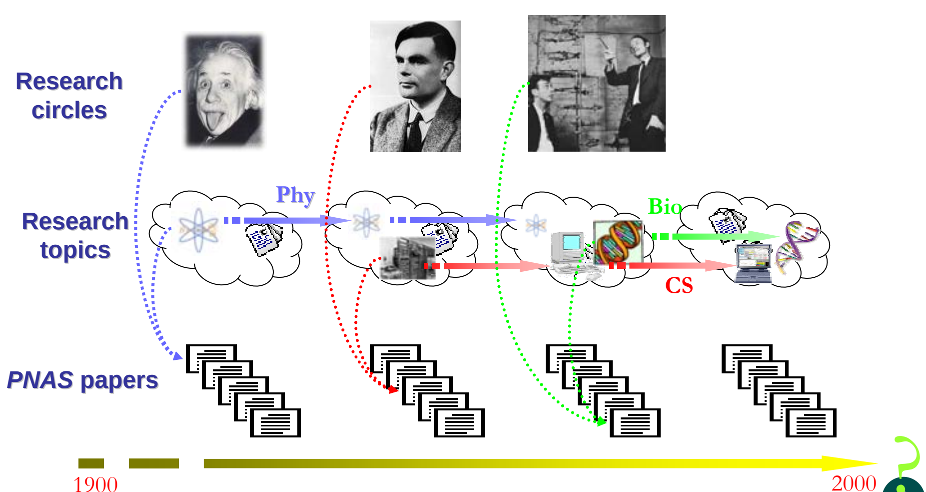
Learn and Recognize Natural Scene Categories



Object Recognition and Tracking

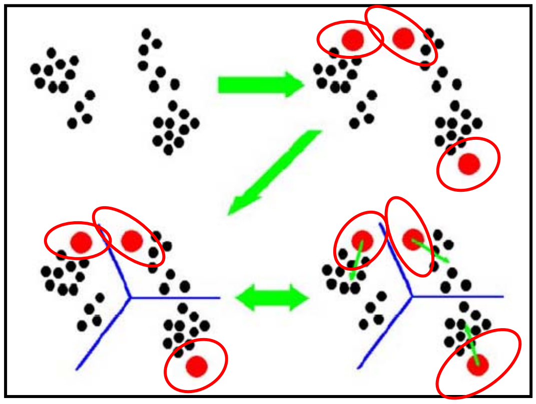


The Evolution of Science



A Classical Approach

- Clustering as Mixture Modeling

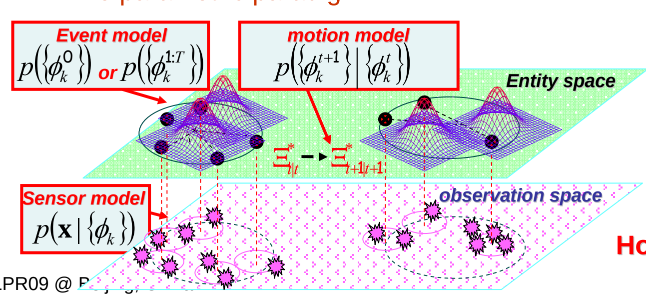


- Then "model selection"

Partially Observed, Open and Evolving Possible Worlds

- Unbounded # of objects/trajectories
- Changing attributes
- Birth/death, merge/split
- Relational ambiguity

- The parametric paradigm:



- Finite
- Structurally unambiguous

How to open it up?

Model Selection vs. Posterior Inference

- Model selection
  - "intelligent" guess: ???
  - cross validation: data-hungry ☹
  - information theoretic:
    - AIC
    - TIC
    - MDL :

    $\arg\min KL\left(f(\cdot) \,|\, g(\cdot \,|\, \hat{\theta}_{ML}, K)\right)$

    Parsimony, Ockam's Razor
  - Bayes factor: need to compute data likelihood

- Posterior inference:

  we want to handle uncertainty of model complexity explicitly

$$p(M \mid D) \propto p(D \mid M) p(M)$$

$$M \equiv \{\theta, K\}$$

  - we favor a distribution that does not constrain M in a "closed" space!

Two "Recent" Developments

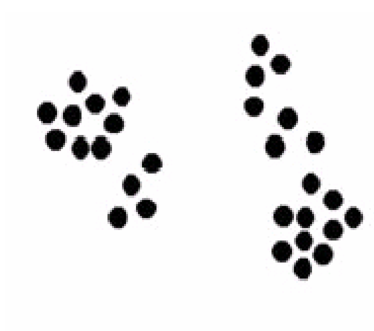
- First order probabilistic languages (FOPLs)
  - Examples: PRM, BLOG …
  - Lift graphical models to "open" world (#rv, relation, index, lifespan …)
  - Focus on complete, consistent, and operating rules to instantiate possible worlds, and formal language of expressing such rules
  - Operational way of defining distributions over possible worlds, via sampling methods

- Bayesian Nonparametrics
  - Examples: Dirichlet processes, stick-breaking processes …
  - From finite, to infinite mixture, to more complex constructions (hierarchies, spatial/temporal sequences, …)
  - Focus on the laws and behaviors of both the generative formalisms and resulting distributions
  - Often offer explicit expression of distributions, and expose the structure of the distributions --- motivate various approximate schemes

Outline

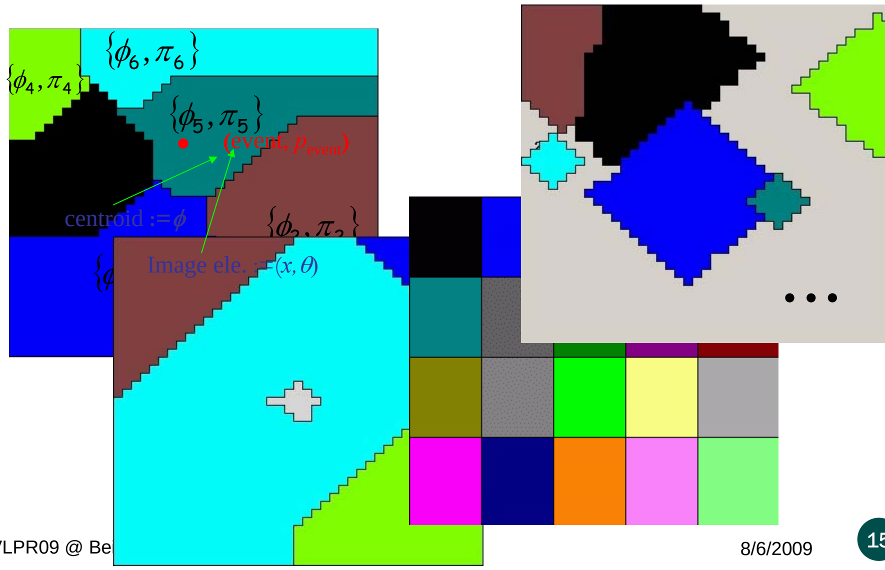
- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Transformed DP and HDP
  - Kernel stick-breaking process
  - Application: joint image segmentation
- Dynamic Dirichlet Process
  - Hidden Markov DP
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary

Clustering



- How to label them ?
- How many clusters ???

Random Partition of Probability Space



Stick-breaking Process

$G \sim \mathrm{DP}(\alpha, G_0)$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k)$$

$$\theta_k \sim G_0$$

Location

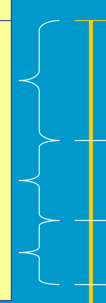
$$\sum_{k=1}^{\infty} \pi_k = 1$$

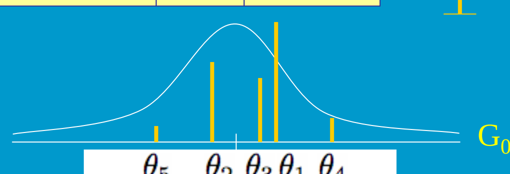
$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_k)$$

Mass

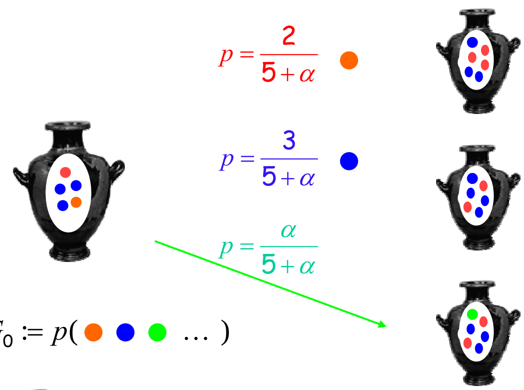
$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

$\prod_{j=1}^{k-1}(1-\beta_j)$	β_k	π_k
0	0.4	0.4
0.6	0.5	0.3
0.3	0.8	0.24





DP – a Pólya urn Process

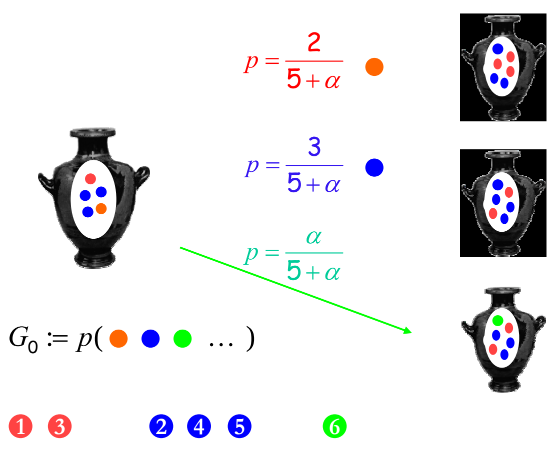


Joint: $G(\;) \sim DP(\alpha G_0)$

Marginal: $\phi_i \mid \phi_{-i}, \alpha, G_0 \sim \sum_{k=1}^{K} \frac{n_k}{i-1+\alpha} \delta_{\phi_k} + \frac{\alpha}{i-1+\alpha} G_0 \,.$

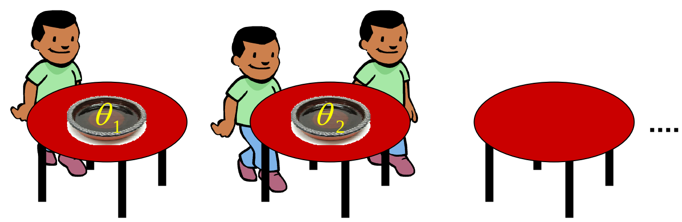
- Self-reinforcing property
- exchangeable partition of samples

Clustering and DP Mixture



- We can associate mixture components with colors in the *Pólya* urn model and thereby define a *clustering* of the data

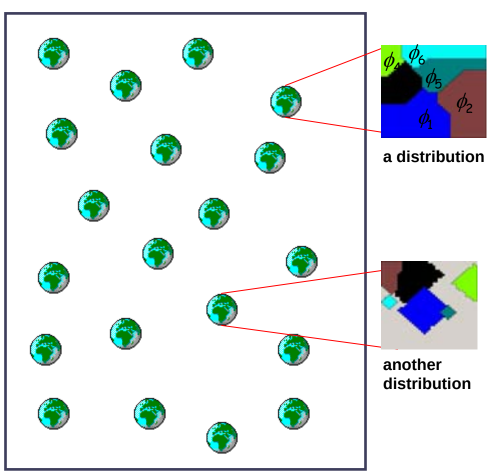
Chinese Restaurant Process



$P(c_i = k \mid \mathbf{c}_{-i}) =$

1	0		0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$		$\frac{\alpha}{2+\alpha}$
$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$		$\frac{\alpha}{3+\alpha}$
$\frac{m_1}{i+\alpha-1}$	$\frac{m_2}{i+\alpha-1}$	....	$\frac{\alpha}{i+\alpha-1}$

Dirichlet Process



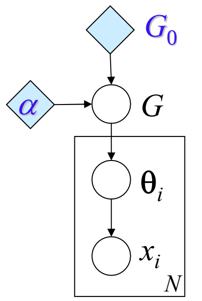
- A *CDF*, G, on possible worlds of random partitions follows a Dirichlet Process if for any measurable finite partition $(\phi_1, \phi_2, .., \phi_m)$:

$(G(\phi_1), G(\phi_2), \ldots, G(\phi_m)) \sim$ Dirichlet$(\alpha G_0(\phi_1), \ldots., \alpha G0(\phi_m))$

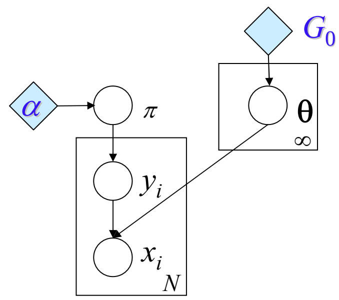
where G_0 is the base measure and α is the scale parameter

Thus a Dirichlet Process G defines a distribution of distribution

Graphical Model Representations of DP



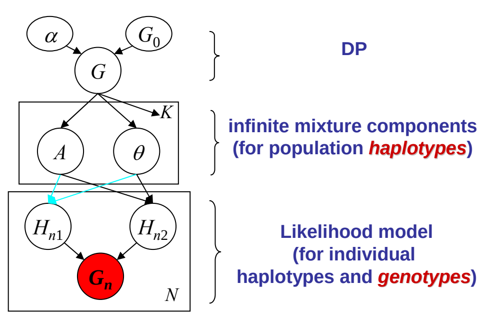
The Pólya urn construction



The Stick-breaking construction

Example: DP-haplotyper [Xing et al, 2004]

- Clustering human populations



- Inference: Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling
  - Metropolis Hasting

Inheritance and Observation Models

- Single-locus mutation model

$$\boxed{A_{C_{i_e}} \rightarrow H_{i_e}}$$

$$P_H(h_t \mid a_t, \theta) = \begin{cases} \theta & \text{for } h_t = a_t \\ \dfrac{1-\theta}{|B|-1} & \text{for } h_t \neq a_t \end{cases}$$

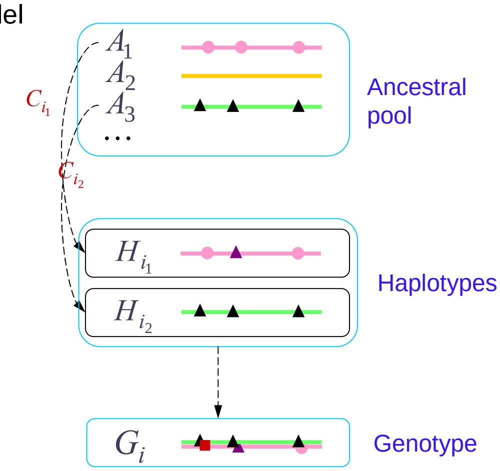
$$\rightarrow h_t = a_t \quad with \quad prob.\ \theta$$

- Noisy observation model

$$\boxed{H_{i_1}, H_{i_2} \rightarrow G_i}$$

$$P_G(g \mid h_1, h_2):$$

$$g_t = h_{1,t} \oplus h_{2,t} \quad with \quad prob.\ \lambda$$



MCMC for Haplotype Inference

- Gibbs sampling for exploring the posterior distribution under the proposed model
  - Integrate out the parameters such as θ or λ, and sample c_{i_e}, a_k and h_{i_e}

$$\underbrace{p(c_{i_e} = k \mid \mathbf{c}_{[-i_e]}, \mathbf{h}, \mathbf{a})}_{\text{Posterior}} \propto \underbrace{p(c_{i_e} = k \mid \mathbf{c}_{[-i_e]})}_{\text{Prior}} \times \underbrace{p(h_{i_e} \mid a_k, \mathbf{h}_{[-i_e]}, \mathbf{c})}_{\text{Likelihood}}$$

Pólya urn

⋮

  - Gibbs sampling algorithm: draw samples of each random variable to be sampled given values of all the remaining variables

MCMC for Haplotype Inference

1. Sample $c_{ie}^{(j)}$, from $p(c_{i_e}^{(j)} = k | \mathbf{c}^{[-j,i_e]}, \mathbf{h}, \mathbf{a})$

$$\propto p(c_{i_e}^{(j)} = k | \mathbf{c}^{[-j,i_e]}, \mathbf{m}, \mathbf{n}) p(h_{i_e}^{(j)} | a_k, \mathbf{c}, \mathbf{h}^{[-j,i_e]})$$

$$\propto (m_{jk}^{[-j,i_e]} + \tau\beta_k) p(h_{i_e}^{(j)} | a_k, \mathbf{l}_k^{[-j,i_e]}), \text{ for } k = 1, ..., K+1$$

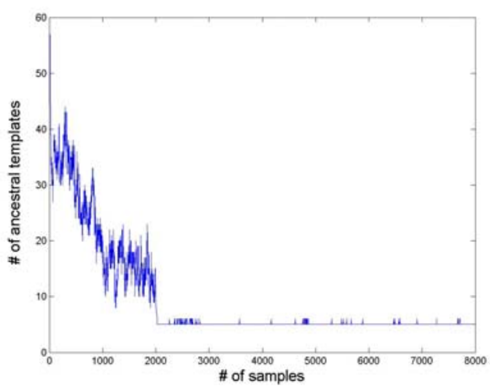
2. Sample a_k from $p(a_{k,t} | \mathbf{c}, \mathbf{h}) \propto \prod_{j,i_e | c_{i_e,t}^{(j)} = k} p(h_{i_e,t}^{(j)} | a_{k,t}, l_{k,t}^{(j)})$

$$= \frac{\Gamma(\alpha_h + l_{k,t})\Gamma(\beta_h + l'_{k,t})}{\Gamma(\alpha_h + \beta_h + m_k)(|B| - 1)^{l'_{k,t}}} R(\alpha_h, \beta_h)$$

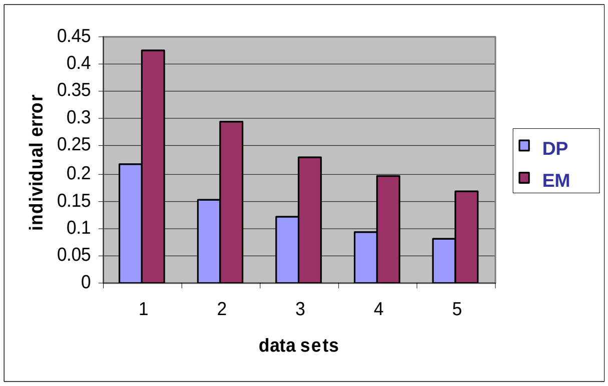
3. Sample $h_{ie}^{(j)}$ from $p(h_{i_e,t}^{(j)} | \mathbf{h}_{[-i_e,t]}^{(j)}, \mathbf{c}, \mathbf{a}, \mathbf{g})$

- For DP scale parameter α: a vague inverse Gamma prior

Convergence of Ancestral Inference



DP vs. Finite Mixture via EM



Variational Inference [Blei & Jordan 2005, Kurihara et al 2007]

- Gibbs sampling solution is not efficient enough to scale up to the large scale problems.
- Truncated stick-breaking approximation can be formulated in the space of explicit, non-exchangeable cluster labels.
- Variational inference can now be applied to such a finite-dimensional distribution

- Variational Inference:
  - For a complicated $P(X_1, X_2, \dots X_n)$, approximate it with $Q(X)$:

$$Q(\mathbf{X}) = \prod_i Q(\mathbf{X}_{C_i})$$

$$\{Q^*(\mathbf{X}_{C_i})\} = \arg\min KL(Q(\mathbf{X})|P(\mathbf{X}))$$

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Approximations to DP

- Truncated stick-breaking representation

$$
\begin{aligned}
v_i &\sim \mathcal{B}(v_i; 1, \alpha) && i = 1, ..., T-1 \\
v_T &= 1 \\
\pi_i &= v_i \prod_{j<i}(1 - v_j) && i = 1, ..., T \\
\pi_i &= 0 && i > T
\end{aligned}
$$

- The joint distribution can be expressed as:

$$
P(X, \mathbf{z}, \mathbf{v}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi}(\mathbf{v}))\right]\left[\prod_{i=1}^{T} p(\eta_i)\mathcal{B}(v_i; 1, \alpha)\right]
$$

- Finite symmetric Dirichlet approximation

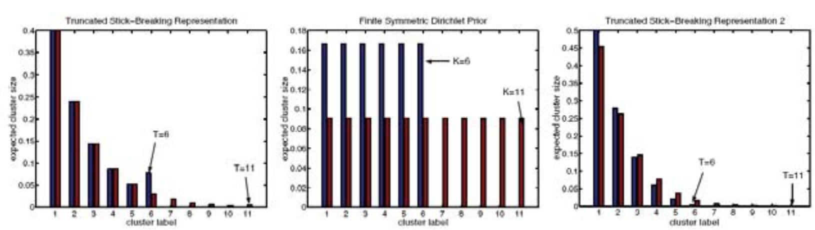
$$
\boldsymbol{\pi} \sim \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, ..., \tfrac{\alpha}{K})
$$

- The joint distribution can be expressed as:

$$
P(X, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi})\right]\left[\prod_{i=1}^{K} p(\eta_i)\right] \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, ..., \tfrac{\alpha}{K})
$$

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

TDP vs. TSB



- TDP is size biased
- cluster labels is NOT interchangeable under TDP but is interchangeable under TSB

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Marginalization

$$P(X, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi})\right] \left[\prod_{i=1}^{K} p(\eta_i)\right] \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, \dots, \tfrac{\alpha}{K})$$

- In variational Bayesian approximation, we assume a factorized form for the posterior distribution.
- However it is not a good assumption since changes in π will have a considerable impact on ***z***.

If we can integrate out π , the joint distribution is given by

$$P(X, \mathbf{z}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\right] p(\mathbf{z}) \left[\prod_{i=1}^{\infty} p(\eta_i)\right]$$

For the TSB representation:

$$p_{\text{TSB}}(\mathbf{z}) = \prod_{i<T} \frac{\Gamma(1+N_i)\Gamma(\alpha+N_{>i})}{\Gamma(1+\alpha+N_{\geq i})} \;\alpha$$

For the FSD representation:

$$p_{\text{FSD}}(\mathbf{z}) = \frac{\Gamma(\alpha)\prod_{k=1}^{K}\Gamma(N_k+\frac{\alpha}{K})}{\Gamma(N+\alpha)\Gamma(\frac{\alpha}{K})^K}$$

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

VB inference

- We can then apply the VB inference on the four approximations

$$\{Q^*(\mathbf{X}_{C_i})\} = \arg\min KL(Q(\mathbf{X})|P(\mathbf{X}))$$

The approximated posterior distribution for TSB and FSD are

$$Q_{\text{TSB}}(\mathbf{z}, \boldsymbol{\eta}, \mathbf{v}) = \left[\prod_{n}^{N} q(z_n)\right] \left[\prod_{i=1}^{T} q(\eta_i) q(v_i)\right]$$

$$Q_{\text{FSD}}(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\pi}) = \left[\prod_{n}^{N} q(z_n)\right] \left[\prod_{k=1}^{K} q(\eta_k)\right] q(\boldsymbol{\pi})$$

Depending on marginalization or not, ***v*** and π may be integrated out.

Experimental results

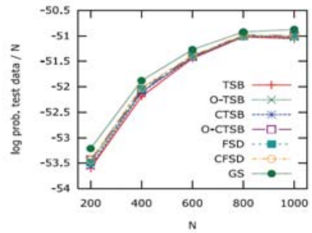


Figure 2: Average log probability per data-point for test data as a function of N.

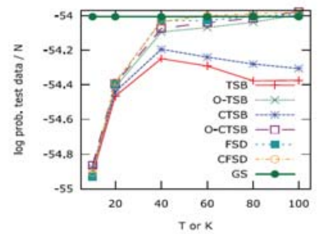
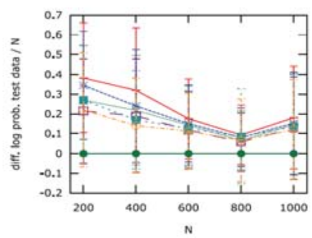
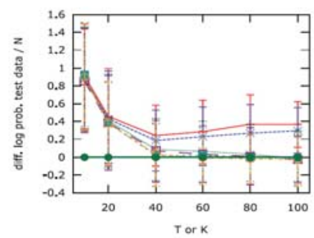


Figure 4: Average log probability per data-point for test data as a function of T (for TSB methods) or K (for FSD methods).

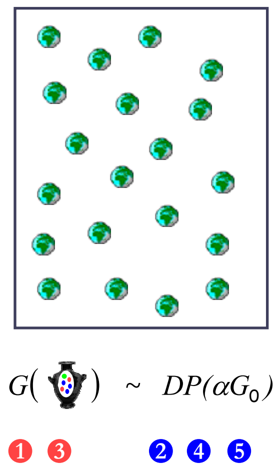


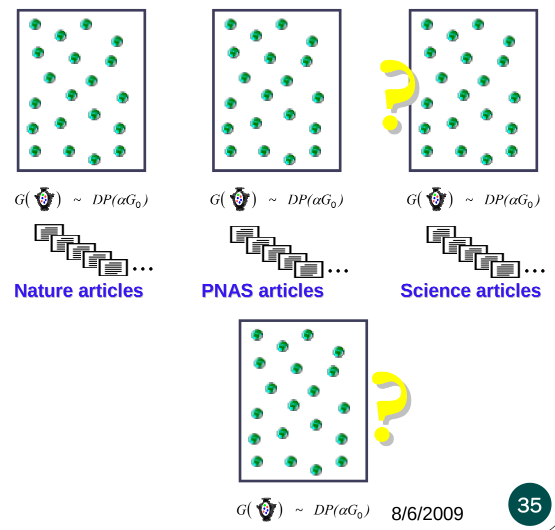


Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Transformed DP and HDP
  - Kernel stick-breaking process
  - Application: joint image segmentation
- Dynamic Dirichlet Process
  - Hidden Markov DP
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary

Solving Multiple Clustering Problems

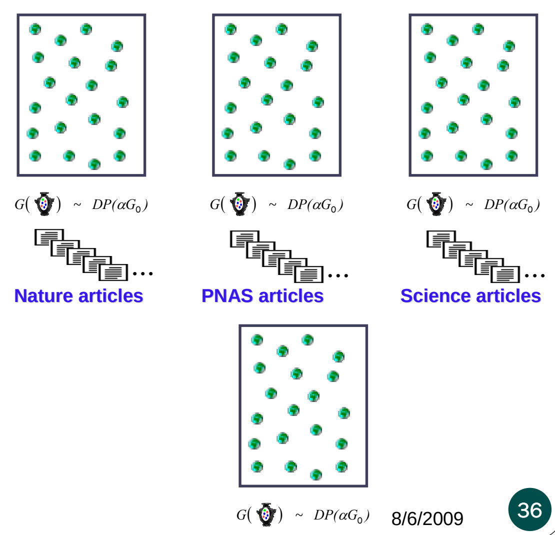




SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegreative Genomics

Solving Multiple Clustering Problems

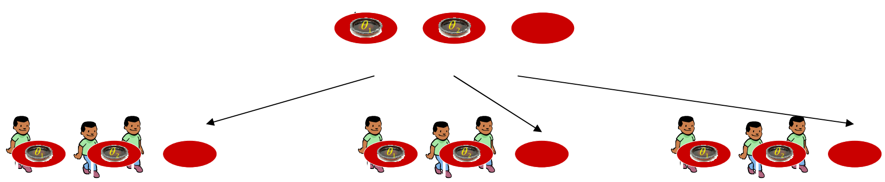
- Solve separately
  - Fail to capture correlation
  - Fail to cross-reinforce shared information
    (i.e., topic specific lexicon)
  - Data fragmentation
- Solve together
  - Then what is the difference between all these journals?



Hierarchical Dirichlet Process

[Teh et al., 2005, Xing et al. 2005]

- **Two level Pólya urn scheme**
  - At the *i*-th step in *j*-th "group",



Oracle

$$-\, Choose\ \theta_k\ with\ prob.\ \frac{m_{jk}}{\sum_k m_{jk} + \alpha_0}$$

$$-\, Go\ to\ the\ upper\ level\ DP\ \ with\ prob.\ \frac{\alpha_0}{\sum_k m_{jk} + \alpha_0}$$

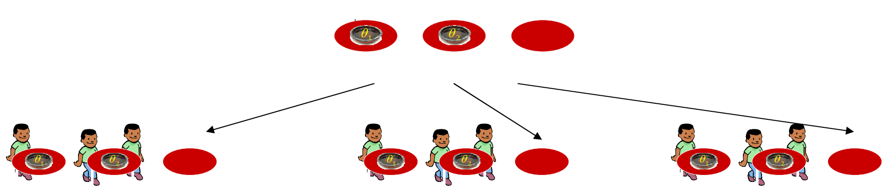
$$Choose\ \theta_k\ with\ prob.\ \frac{n_k}{\sum n_k + \gamma}$$

$$Draw\ a\ new\ sample\ \ with\ prob.\ \frac{\gamma}{\sum n_k + \gamma}$$

Hierarchical Dirichlet Process

[Teh et al., 2005, Xing et al. 2005]

- **Two level Pólya urn scheme**
  - At the *i*-th step in *j*-th "group",



- Draw from stock urn define Dirichlet Process $DP(\gamma,H)$

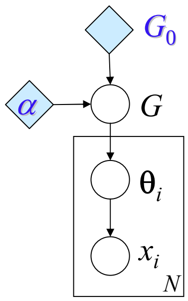
$$\theta_i \mid \theta_{-i} \sim \sum_{k=1}^{K} \frac{n_k}{i+\gamma} \delta_{\phi_k^*}(\theta_i) + \frac{\gamma}{i+\gamma} H(\theta_i)$$

- Conditioning on $DP(\gamma,H)$, the m_jth draw from the mth bottom-level urn also form a Dirichlet measure

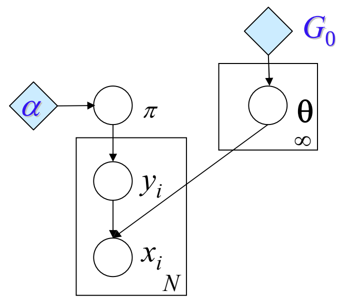
$$\theta_{m_j} \mid \theta_{-m_j} \sim \sum_{k=1}^{K} \frac{m_{j,k} + \alpha \frac{n_k}{n+\gamma}}{m_j + \alpha} \delta_{\phi_k^*}(\theta_{m_j}) + \frac{\alpha}{m_j+\alpha} \frac{\gamma}{i+\gamma} H(\theta_{m_j})$$

$$= \sum_{k=1}^{K} p_k^* \delta_{\phi_k^*}(\theta_{m_j}) + p_{K+1}^* H(\theta_{m_j})$$

Recall: Graphical Model Representations of DP

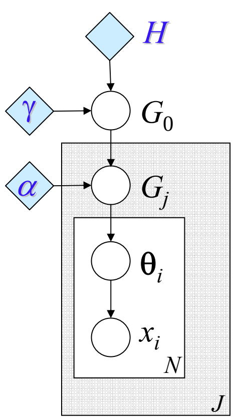


The Pólya urn construction



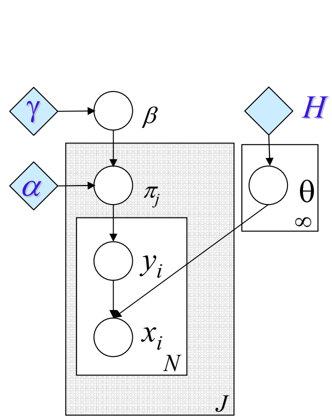
The Stick-breaking construction

Hierarchical DP Mixture



Stick(α, β):

$\pi'_{jk} \sim \text{Beta}\left(\alpha\beta_k, \alpha\left(1-\sum_{l=1}^{k}\beta_l\right)\right), \quad \pi_{jk} = \pi'_{jk}\prod_{l=1}^{k-1}\left(1-\pi'_{jl}\right).$



$\theta_k \sim H$

$\beta = \text{Stick}(\gamma), G_0 = \sum_{k=1}^{\infty}\beta_k\delta(\theta_k)$

$\pi_j = \text{Stick}(\alpha, \beta), G_j = \sum_{k=1}^{\infty}\pi_k\delta(\theta_k)$

Topic Models for Images

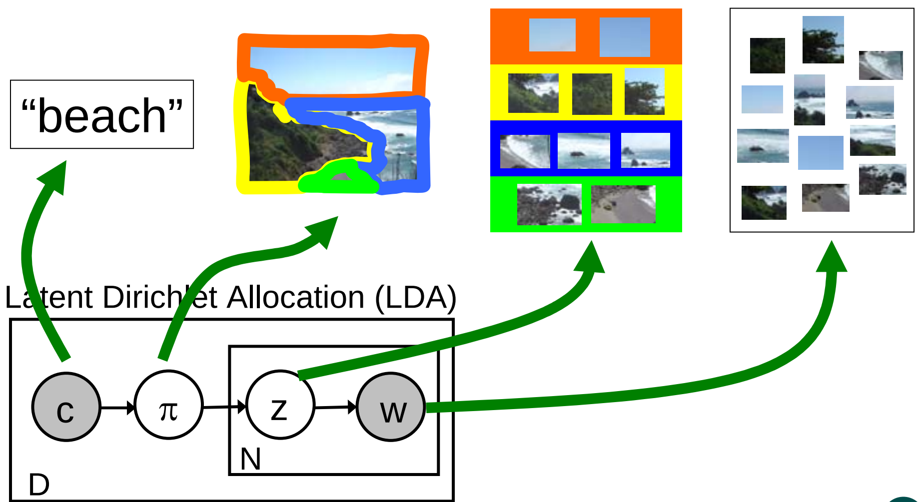
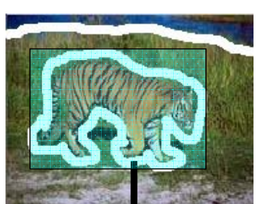


Image Representation



cat, grass, tiger, water

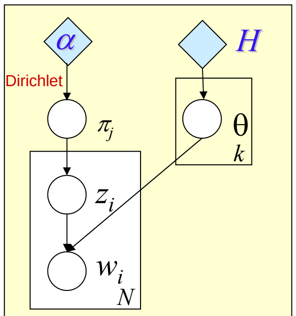
$[r_{11} \cdots r_{1d}]$, $[w_1 \cdots w_{|V|}]$

representation vector (real, 1 per image segment) : **annotation vector** (binary, same for each segment)

$$[r_{n1} \cdots r_{nd}] \ , \ [w_1 \cdots w_{|V|}]$$

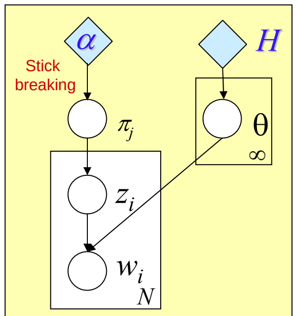
SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Infinite Topic Model for Image



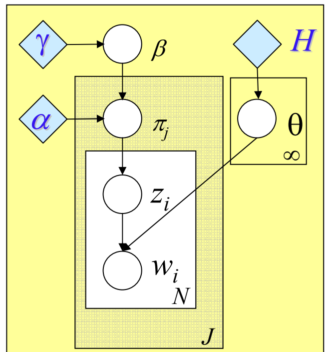
A single image with *k* topic

An LDA



A single image with inf-topic

A DP



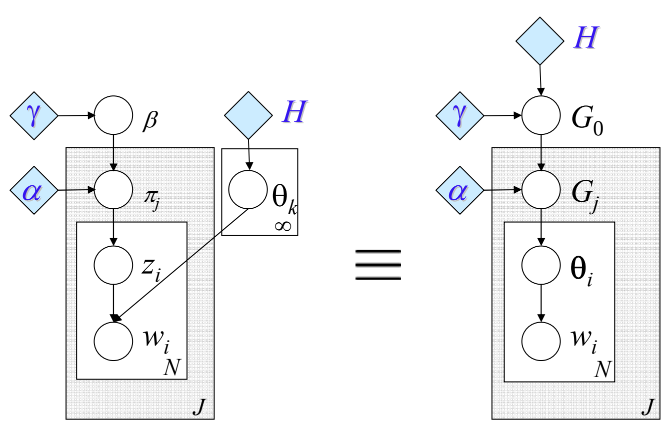
***J* images with inf-topic**

An HDP

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Problem with HDP

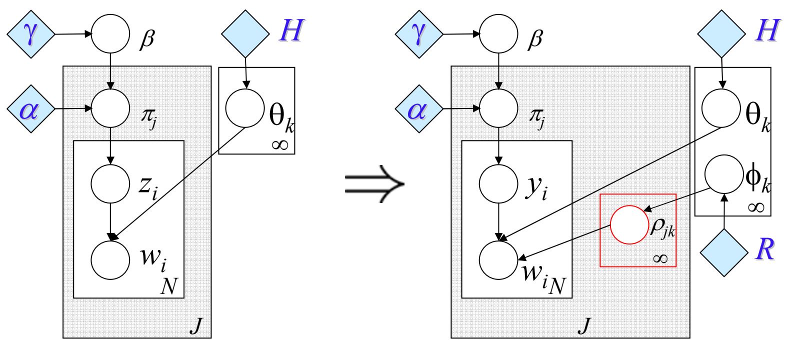
- Every group (i.e., image) has exactly the same set of visual-vocabulary topics, albeit with different frequency



Transformed Dirichlet Process

[Sudderth et al, 2005]

- An extension of HDP in which global mixture components undergo a set of random transformations before being reused in each group



Synthetic Data Results

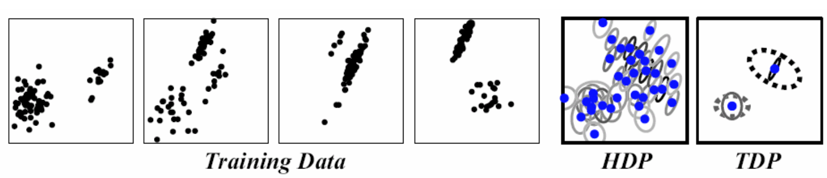
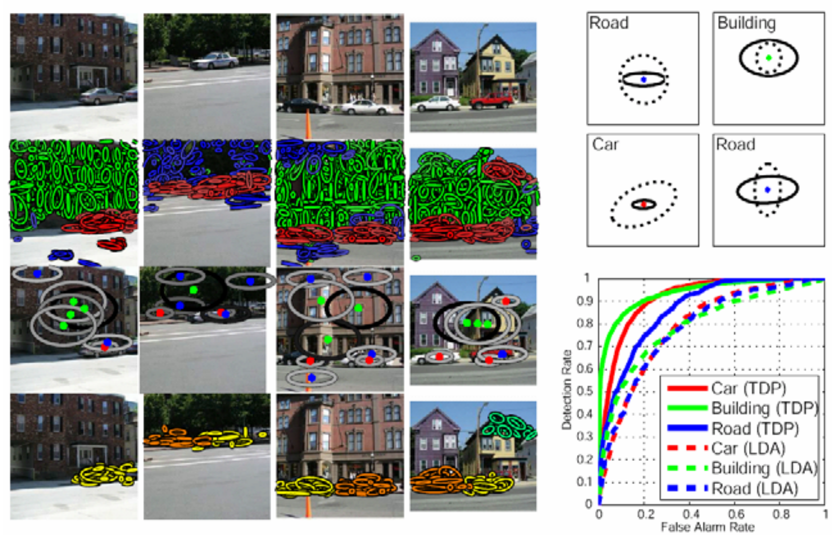


Figure 3: Comparison of hierarchical models learned via Gibbs sampling from synthetic 2D data. *Left:* Four of 50 "images" used for training. *Center:* Global distribution $G_0(\theta)$ for the HDP, where ellipses are covariance estimates and intensity is proportional to prior probability. *Right:* Global TDP distribution $G_0(\theta, \rho)$ over both clusters θ (solid) and translations ρ of those clusters (dashed).

- HDP uses a large set of global clusters to discretize the transformations underlying the data, and may have poor generalization for modeling visual scenes.

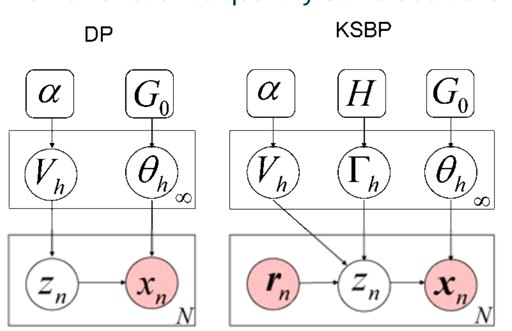
Analyzing Street Scenes



Kernel stick-breaking process

[Dunson and Park, 2006]

- For image analysis, we want to impose the belief that spatially proximate patches are more probable to be associated with the same cluster.
- We augmented the stick-breaking representation of DP to employ a kernel function to quantify some additional prior.



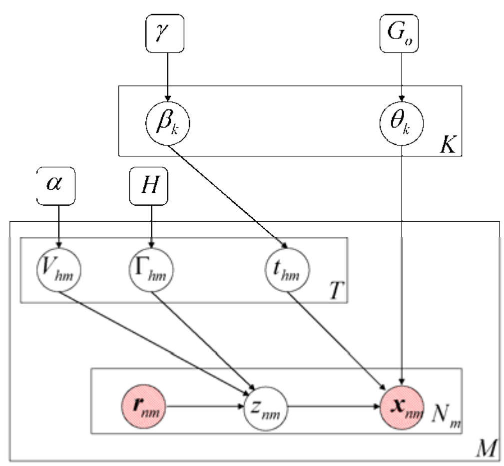
KSBP for image analysis

N

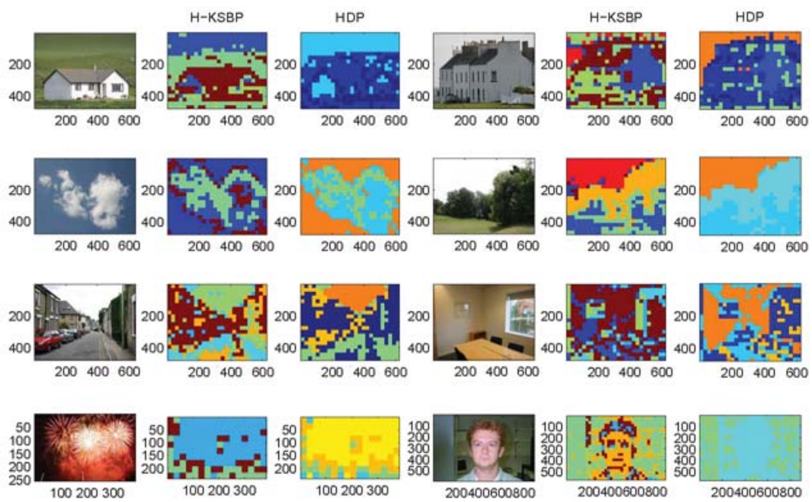
- Consider a image composed of N patches, the features vectors $\{\boldsymbol{x}_n\}_{n=1,N}$ and the associated locations $\{\boldsymbol{r}_n\}_{n=1,N}$ can be modeled as follows:

$$
\begin{aligned}
\boldsymbol{x}_n &\overset{ind}{\sim} f(\phi_n) \\
\phi_n &\overset{ind}{\sim} G_r \\
G_r &= \sum_{h=1}^{\infty} \pi_h(\mathbf{r}; V_h, \Gamma_h, \psi)\delta_{\theta_h} \\
\pi_h(\mathbf{r}; V_h, \Gamma_h, \psi) &= V_h K(\mathbf{r}, \Gamma_h, \psi) \prod_{l=1}^{h-1} [1 - V_l K(\mathbf{r}, \Gamma_l, \psi)] \\
V_h &\overset{iid}{\sim} Beta(a, b) \\
\Gamma_h &\overset{iid}{\sim} H \\
\theta_h &\overset{iid}{\sim} G_o
\end{aligned}
$$

Multi-task Image Segmentation



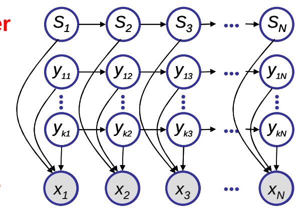
Segmentation results [An et al, 2008]



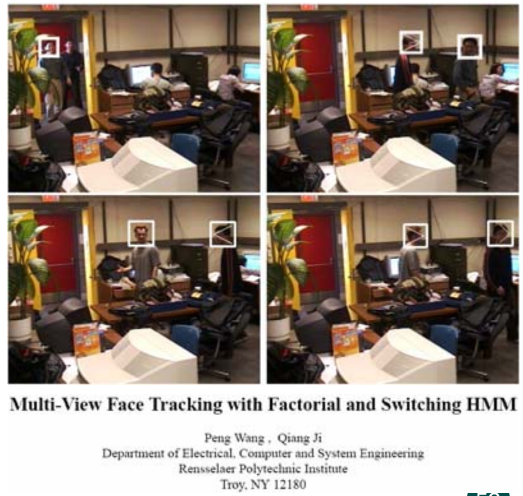
Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Transformed DP and HDP
  - Kernel stick-breaking process
  - Application: joint image segmentation
- Dynamic Dirichlet Process
  - Hidden Markov DP
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary

Object Recognition and Tracking



- Each chain corresponds to the trajectory of a specific object

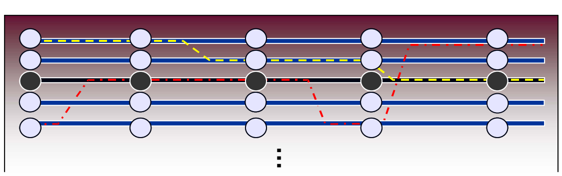


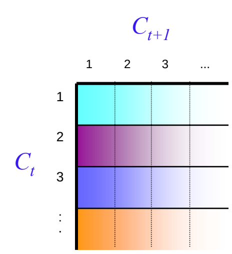
Hidden Markov Dirichlet Process

(Xing and Sohn. Bayesian Analysis, 2007, Sohn and Xing, ISMB 2007)

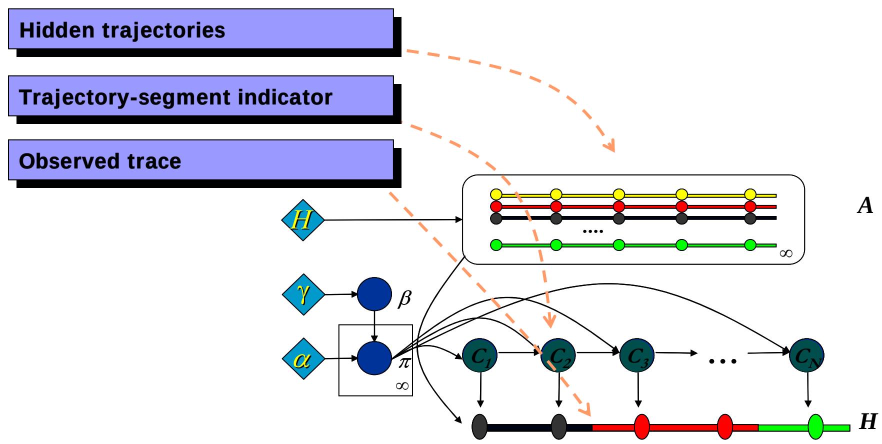
- Hidden Markov Dirichlet process mixtures
  - Extension of HMM model to infinite ancestral space
    - Infinite dimensional transition matrix
    - Each row of the transition matrix is modeled with a DP: $G_i \mid \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$

$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$



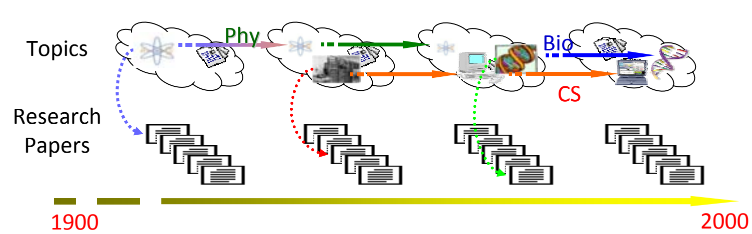


HMDP as a Graphical Model



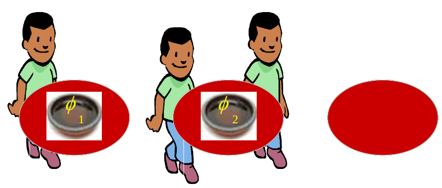
Evolutionary Clustering

- Adapts the number of mixture components over time
  - Mixture components can die out
  - New mixture components are born at any time
  - Retained mixture components parameters evolve according to a Markovian dynamics



The Chinese Restaurant Process

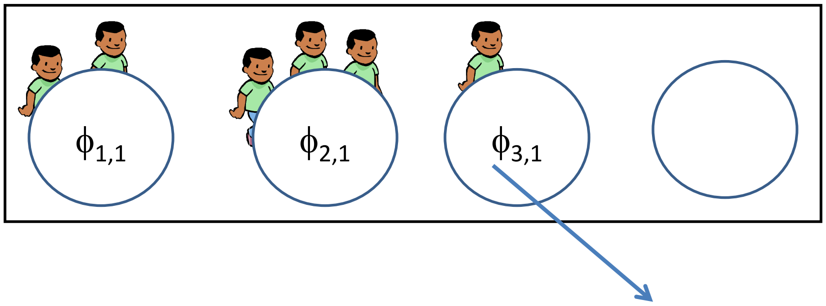
- Customers correspond to data points
- Tables correspond to clusters/mixture components
- Dishes correspond to parameter of the mixtures



Temporal DPM [Ahmed and Xing 2008]

- **The Recurrent Chinese Restaurant Process**
  - The restaurant operates in epochs
  - The restaurant is closed at the end of each epoch
  - The state of the restaurant at time epoch *t* depends on that at time epoch *t-1*
    - Can be extended to higher-order dependencies.

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics



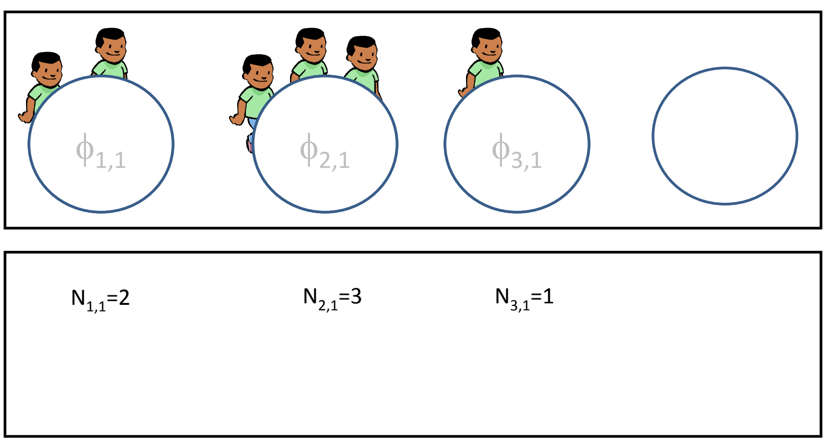
T=1

Dish eaten at table 3 at time epoch 1
OR the parameters of cluster 3 at time epoch 1

Generative Process

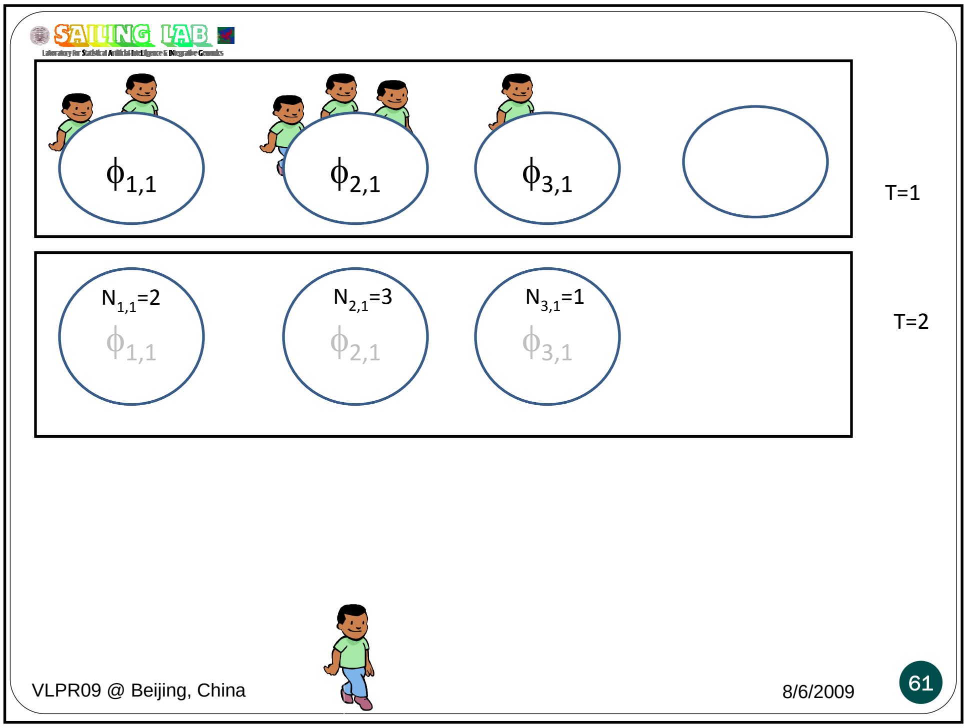
-Customers at time T=1 are seated as before:
- Choose table $j \propto N_{j,1}$ and Sample $x_i \sim f(\phi_{j,1})$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1,1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1,1})$

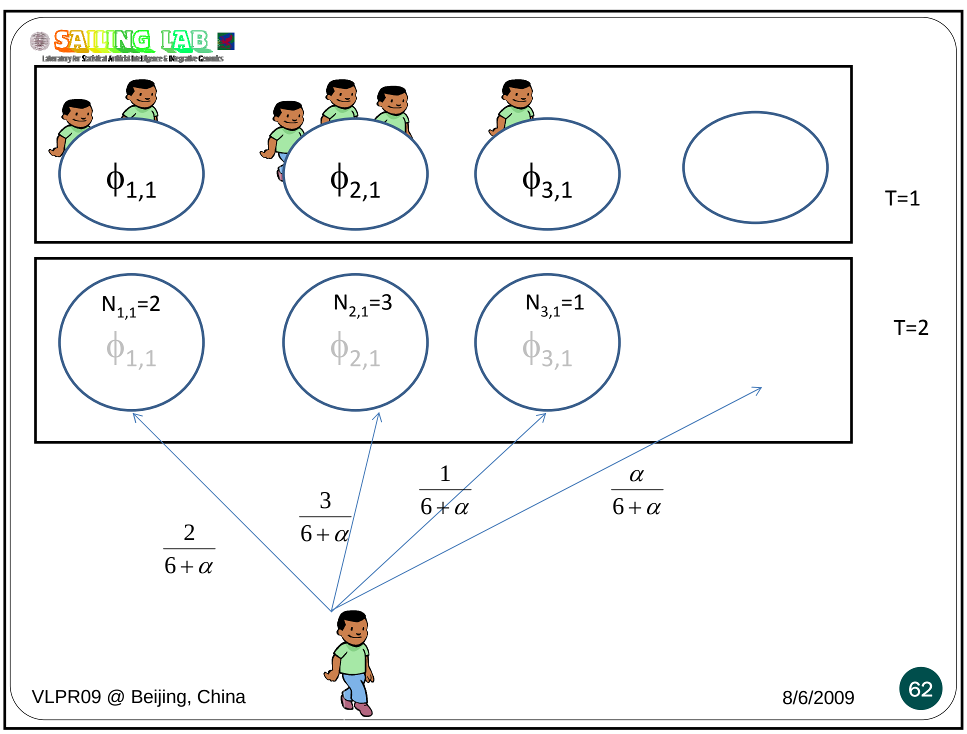
SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

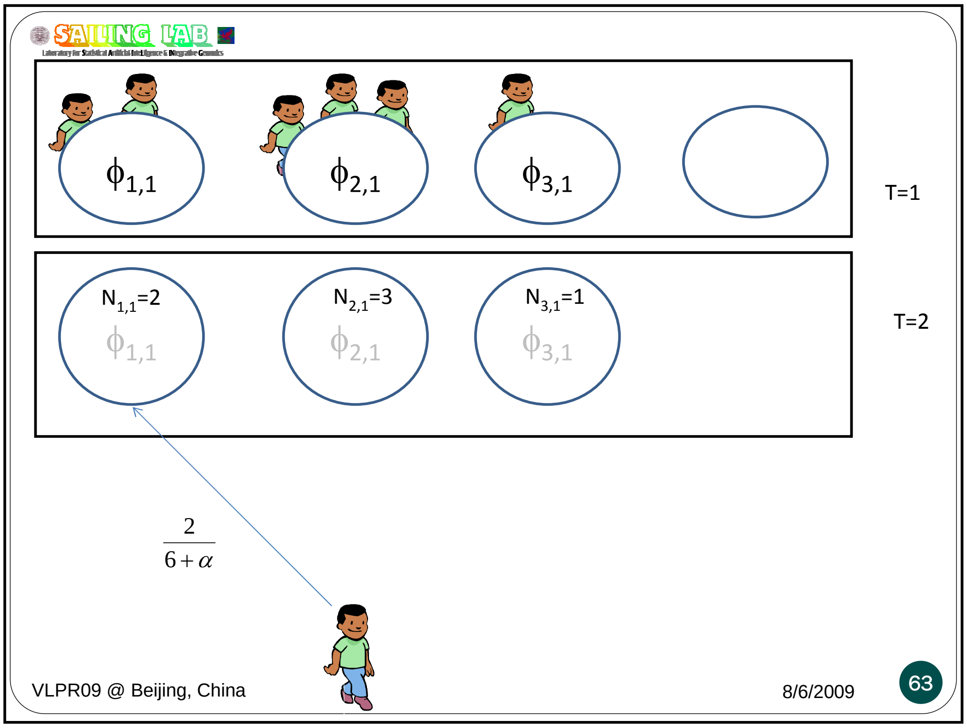


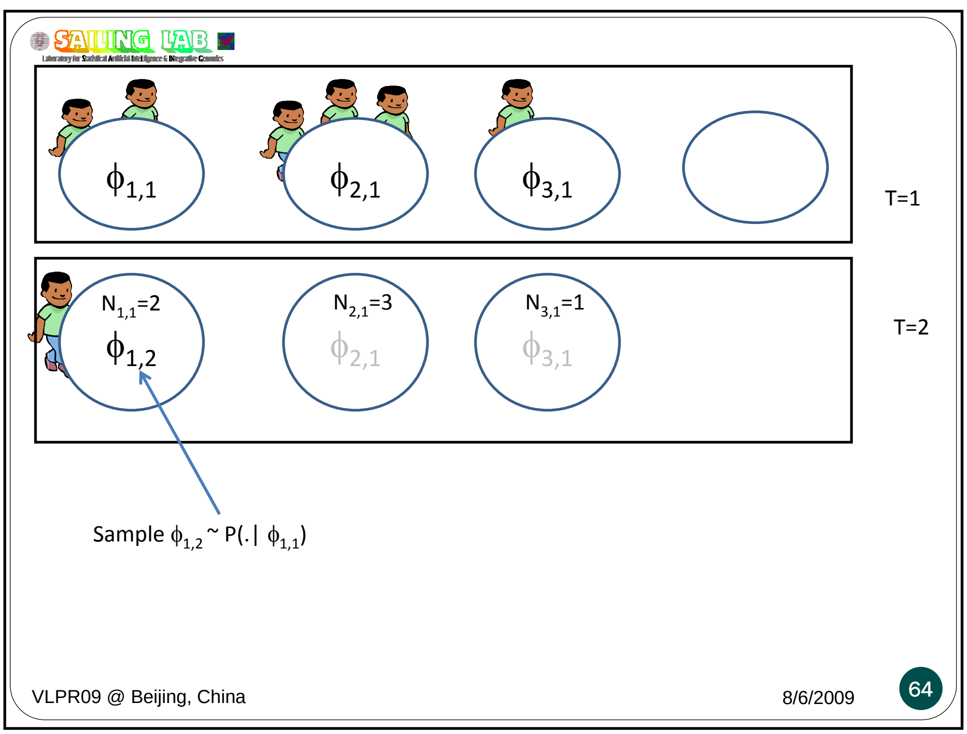
T=1

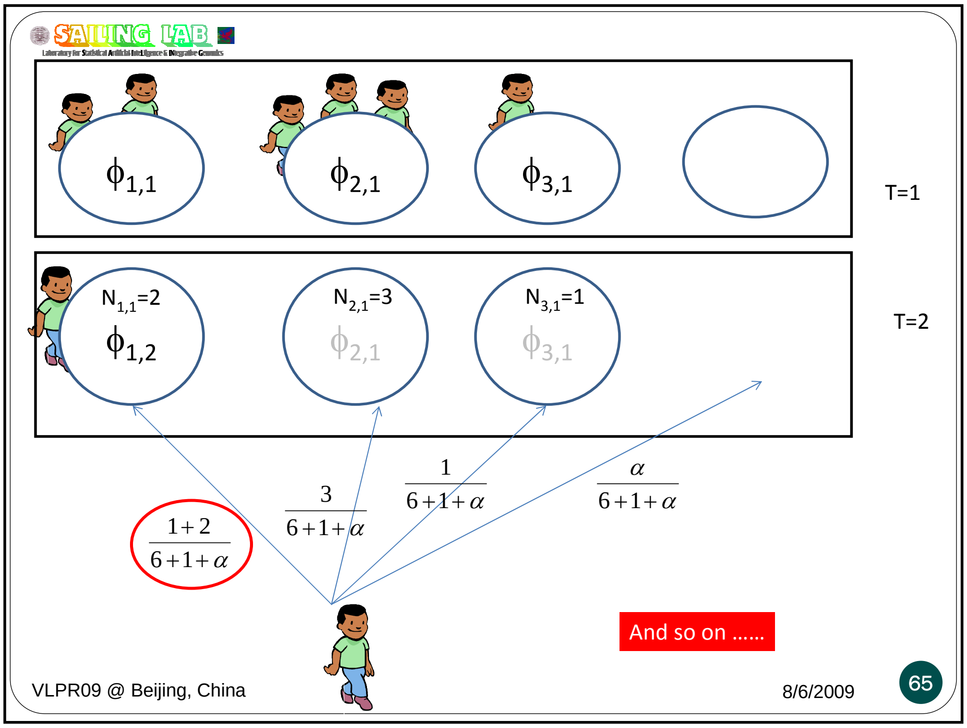
T=2

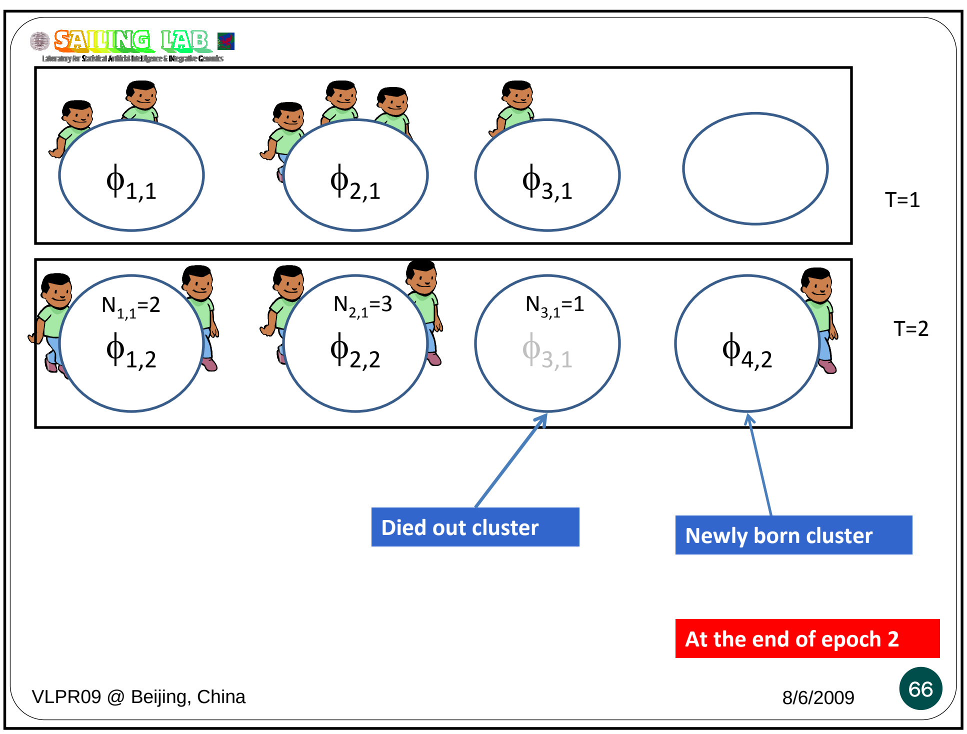

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics
$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$N_{1,1}=2$
$\phi_{1,1}$
$N_{2,1}=3$
$\phi_{2,1}$
$N_{3,1}=1$
$\phi_{3,1}$
T=2
VLPR09 @ Beijing, China
8/6/2009
61

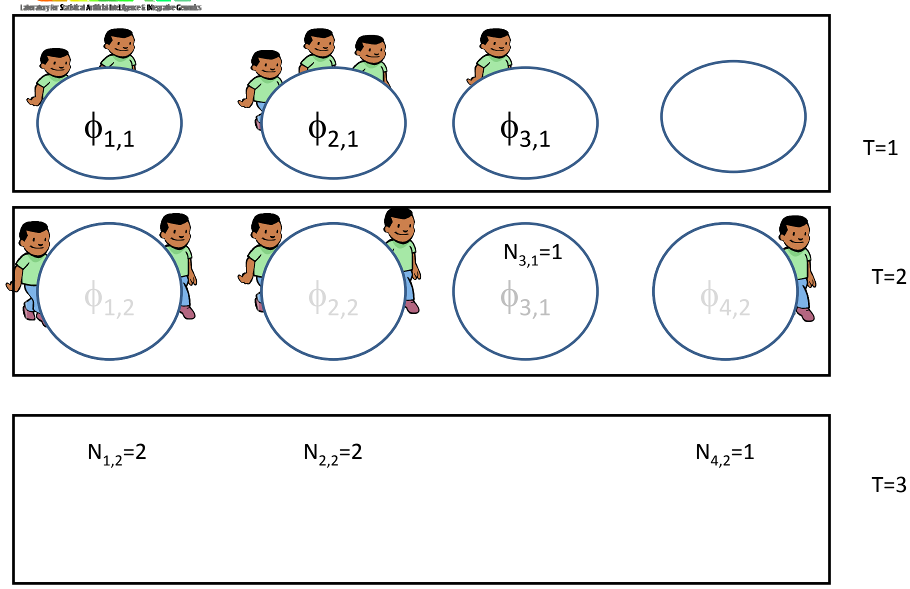

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics
$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$N_{1,1}=2$
$\phi_{1,1}$
$N_{2,1}=3$
$\phi_{2,1}$
$N_{3,1}=1$
$\phi_{3,1}$
T=2
$\frac{2}{6+\alpha}$
$\frac{3}{6+\alpha}$
$\frac{1}{6+\alpha}$
$\frac{\alpha}{6+\alpha}$
VLPR09 @ Beijing, China
8/6/2009
62


SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics
$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$N_{1,1}$=2
$\phi_{1,1}$
$N_{2,1}$=3
$\phi_{2,1}$
$N_{3,1}$=1
$\phi_{3,1}$
T=2
$\frac{2}{6+\alpha}$
VLPR09 @ Beijing, China
8/6/2009
63


SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics
$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$N_{1,1}$=2
$\phi_{1,2}$
$N_{2,1}$=3
$\phi_{2,1}$
$N_{3,1}$=1
$\phi_{3,1}$
T=2
Sample $\phi_{1,2}$ ~ P(.| $\phi_{1,1}$)
VLPR09 @ Beijing, China
8/6/2009
64


SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics
$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$N_{1,1}=2$
$\phi_{1,2}$
$N_{2,1}=3$
$\phi_{2,1}$
$N_{3,1}=1$
$\phi_{3,1}$
T=2
$\frac{1+2}{6+1+\alpha}$
$\frac{3}{6+1+\alpha}$
$\frac{1}{6+1+\alpha}$
$\frac{\alpha}{6+1+\alpha}$
And so on ......
VLPR09 @ Beijing, China
8/6/2009
65


SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics
$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$N_{1,1}=2$
$\phi_{1,2}$
$N_{2,1}=3$
$\phi_{2,2}$
$N_{3,1}=1$
$\phi_{3,1}$
$\phi_{4,2}$
T=2
Died out cluster
Newly born cluster
At the end of epoch 2
VLPR09 @ Beijing, China
8/6/2009
66



Temporal DPM

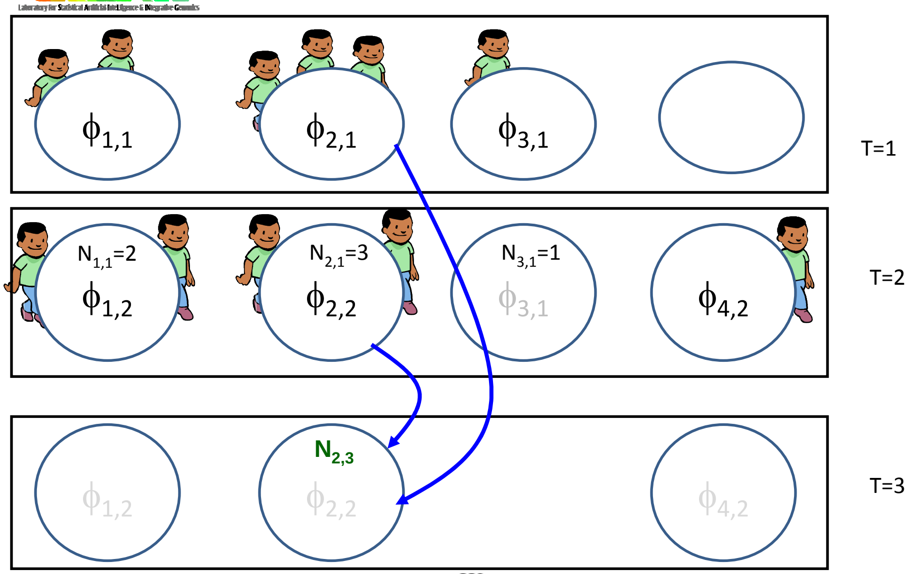
- Can be extended to model higher-order dependencies
- Can decay dependencies over time
  - Pseudo-counts for table k at time t is

$$\sum_{w=1}^{W}\left(\underbrace{e^{\frac{-w}{\lambda}}}_{\text{Decay factory}}\underbrace{N_{k,t-w}}_{\text{Number of customers sitting at table } K \text{ at time epoch } t-w}\right)$$

History size

Decay factory

Number of customers sitting at table *K* at time epoch *t-w*



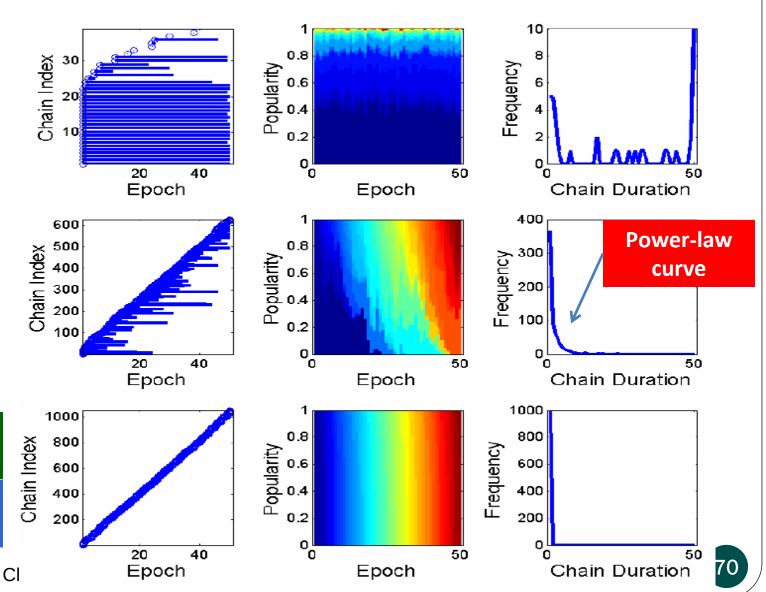
$$N_{2,3} = \sum_{w=1}^{W}\left(e^{\frac{-w}{\lambda}} N_{k,t-w}\right)$$

TDPM Generative Power

DPM
W=T
$\lambda = \infty$

TDPM
W=4
$\lambda = .4$

Independent DPMs
W= 0
λ = ? (any)



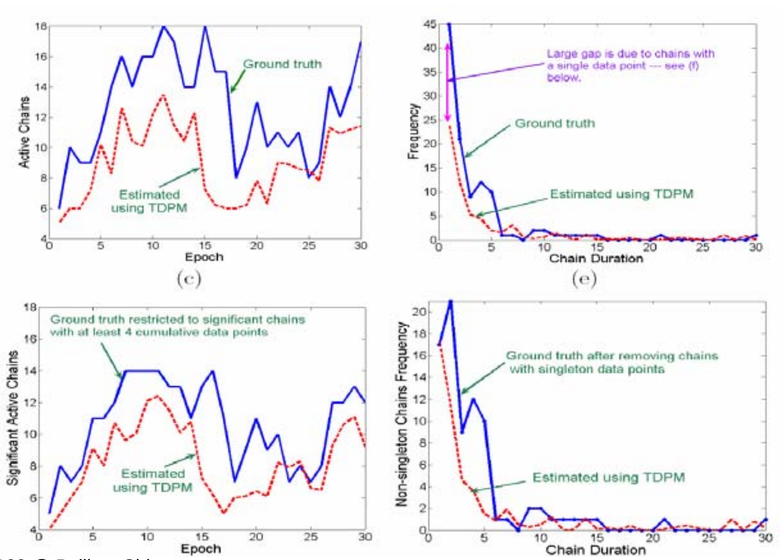
SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Experiments

- Simulated data
- Chain dynamics is modeled as random walk
$$\phi_{k,t}|\phi_{k,t-1} \sim N(\phi_{k,t-1}, \rho I)$$
- Gaussian emission: $x_{t,i}|c_{t,i} = k \sim N(\phi_{k,t}, \Sigma)$
- Simulated 30 epochs with 100 data points in each epoch
- Can TDPM recover the ground truth clustering?
  - Posterior inference ran using Gibbs sampling [Ahmed and Xing 2008]
- Compare with fixed-dimension dynamic models

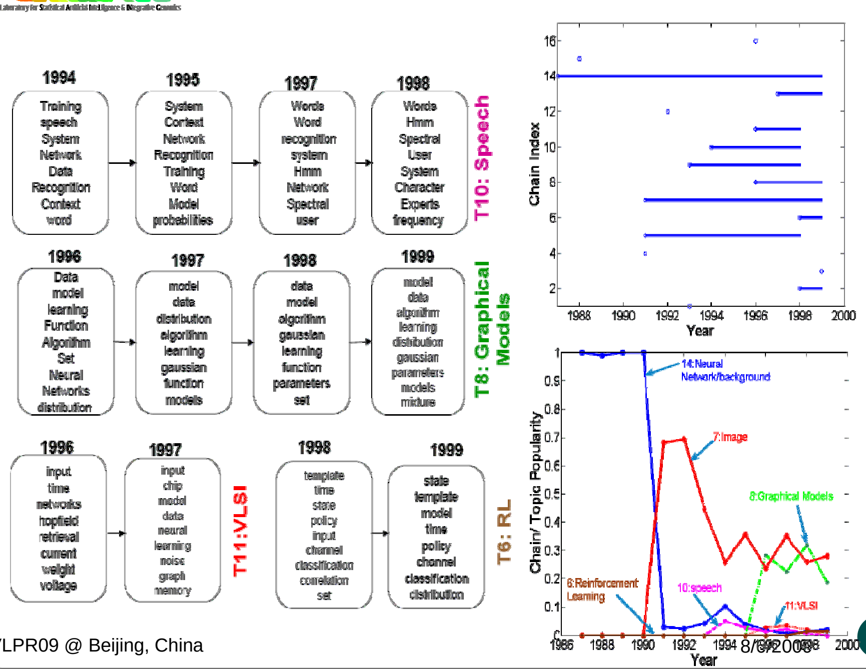
SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

TDPM Adaptability over Time

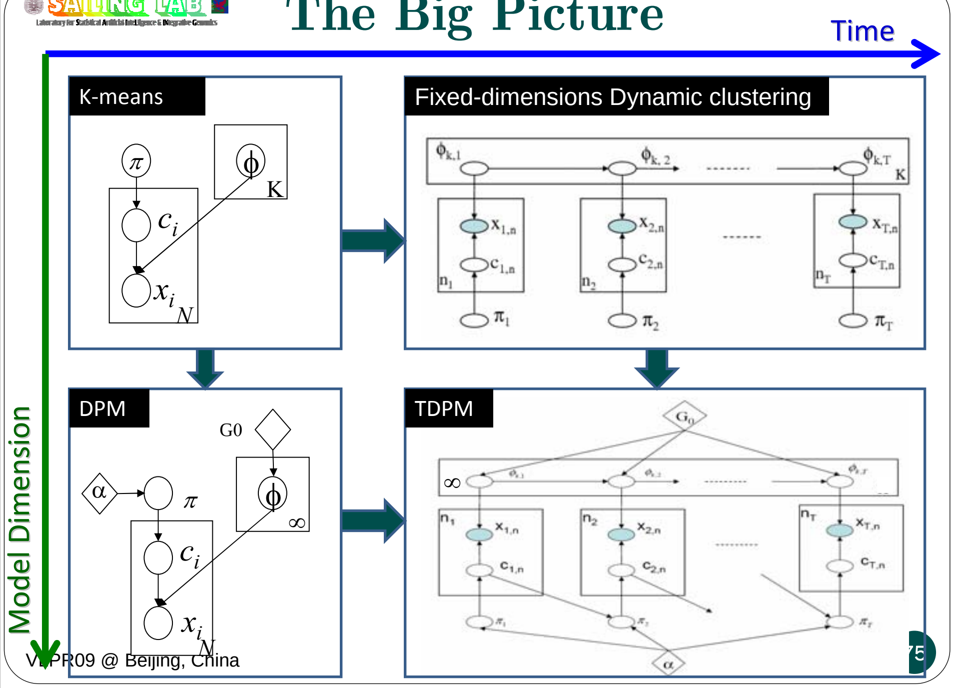


Results: NIPS 12

- Building a simple dynamic topic model
- Chain dynamics is as before
- Emission model for document $x_{k,t}$ is:
  - Project $\phi_{k,t}$ over the simplex
  - Sample $x_{k,t}|c_{t,i} \sim \text{Multinomial}(.|\text{Logisitic}(\phi_{k,t}))$
- Unlike LDA here a document belongs to one topic
- Use this model to analyze NIPS12 corpus
  - Proceeding of NIPS conference 1987-1999



The Big Picture



Summary

- A non-parametric Bayesian model for Pattern Uncovery
  - Finite mixture model of latent patterns (e.g., image segments, objects)
    → infinite mixture of propotypes: alternative to model selection
    → hierarchical infinite mixture
    → infinite hidden Markov model
    → temporal infinite mixture model
- Applications in general data-mining …

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

How to Model Semantic?

- Q: What is it about?
- A: Mainly MT, with syntax, some learning

0.6	0.3	0.1
MT	Syntax	Learning

Mixing Proportion

Source Target SMT Alignment Score BLEU	Parse Tree Noun Phrase Grammar CFG	likelihood EM Hidden Parameters Estimation argMax

Topics

Unigram over vocabulary

A Hierarchical Phrase-Based Model for Statistical Machine Translation

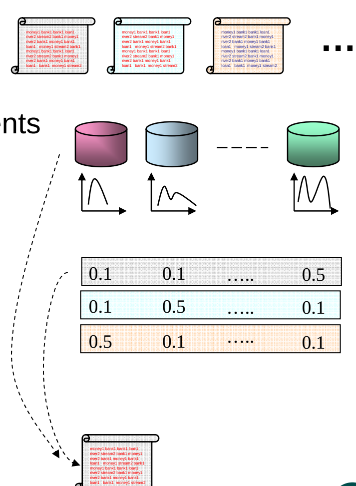
We present a statistical phrase-based Translation model that uses *hierarchical phrases*—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the *formal* machinery of syntax based translation systems without any *linguistic* commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative Improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

Topic Models

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Admixture Models

- Objects are bags of elements
- Mixtures are distributions over elements
- Objects have mixing vector θ
  - Represents each mixtures' contributions
- Object is generated as follows:
  - Pick a mixture component from θ
  - Pick an element from that component



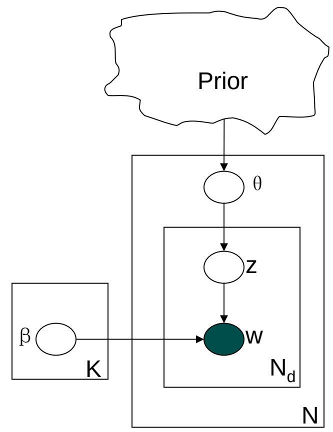
Topic Models = Admixture Models

Generating a document

– *Draw* θ from the prior

For each word n

- Draw z_n from $multinomial(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$



Which prior to use?

Variational Inference