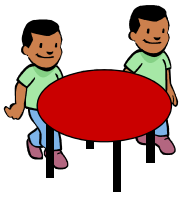
Machine Learning

10-701/15-781, Fall 2011

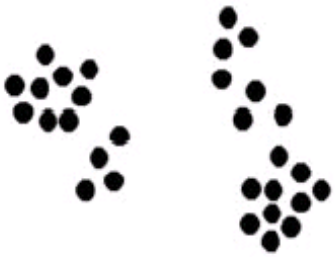
Infinite Mixture Models

Eric Xing

Lecture 10, October 12, 2011



Clustering

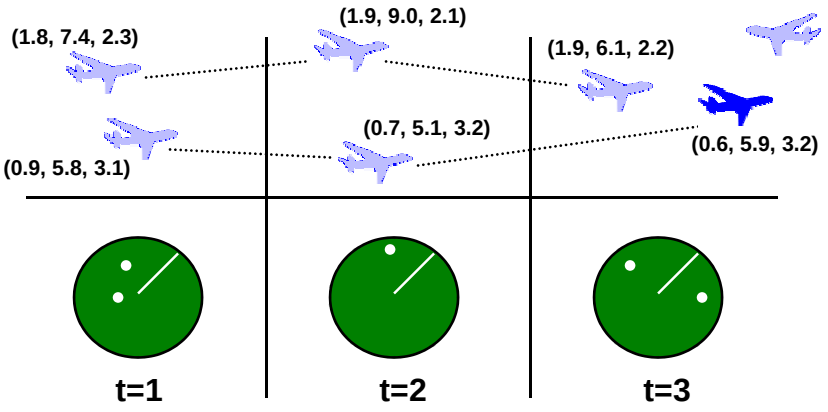


- How to label them ?
- How many clusters ???

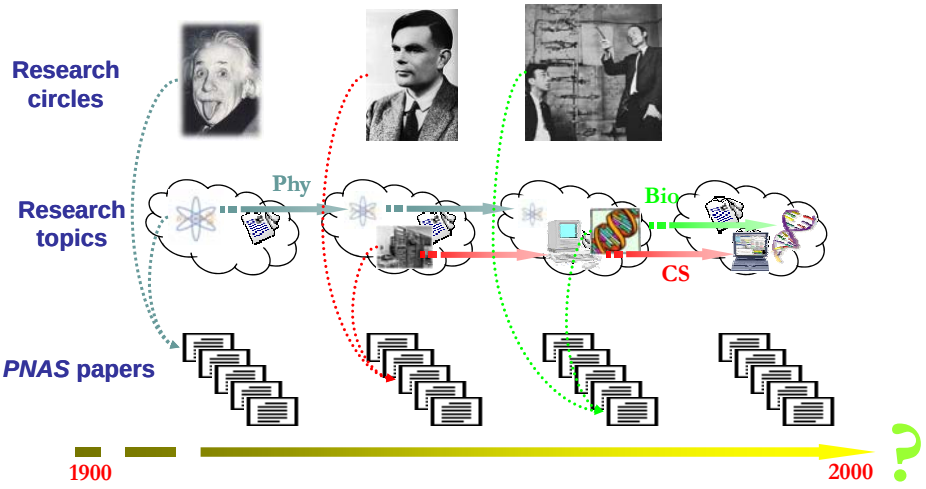
Image Segmentation

- 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*

Object Recognition and Tracking

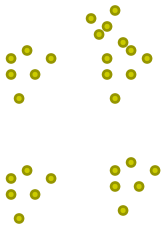


The Evolution of Science



Clustering as (nonparametric) Bayesian Inference

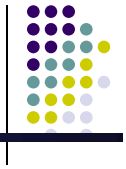
- Cluster = Mixture
- Each data point is generated from one Mixture



How Many clusters

-Cross Validation
  -Data hungry!
-Information theoretic
  -AIC, MDL, etc.
-Non-Parametric
  - Manage model uncertainty
  - Integrate over different clustering configurations
  - Base for many interesting extensions

Outline



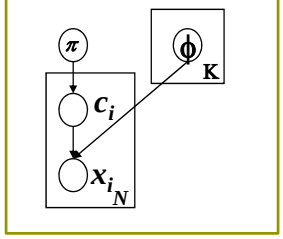
- Dirichlet Process: From finite mixture to Infinite mixture model
  - The Chinese Restaurant Process
  - Example: heliotype inference
- Intro to Markov Chain Monte Carlo
  - Gibbs sampling
- Dynamic Dirichlet Process
  - The recurrent Chinese Restaurant Process
  - Example: Application: evolutionary clustering of documents

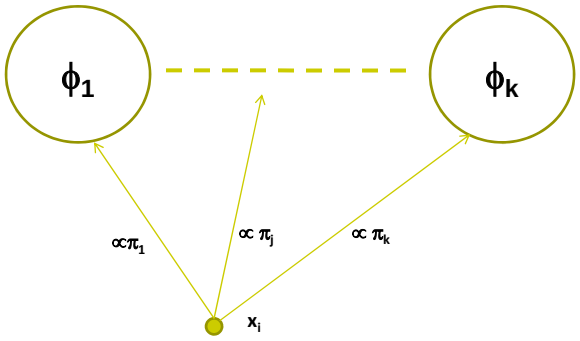
Clustering

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

Mixture model: The Generative Process

Mixture of G





Generative Process

-For data point x_i
- Sample $c_i \sim$ Multi(π)
- Sample $x_i \sim f(\phi_{ci})$

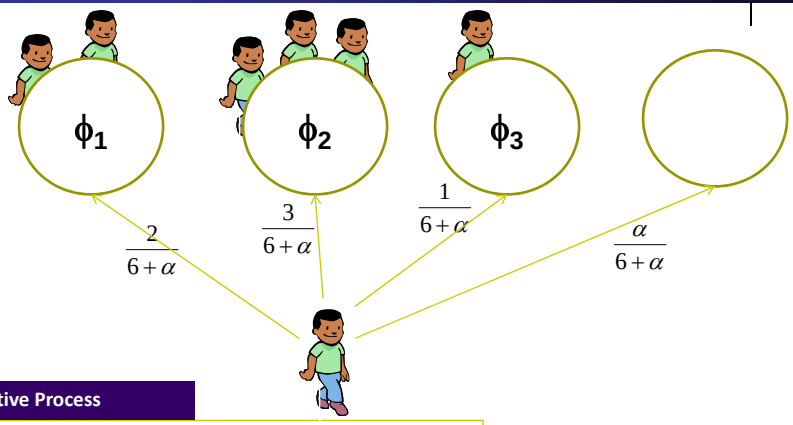
Cluster parameters DO NOT evolve over time

Number of clusters DOES NOT grow with the data

The Chinese Restaurant Process

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

The Chinese Restaurant Process



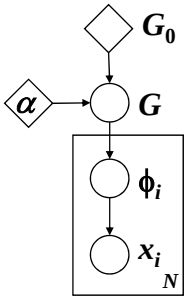
Generative Process

-For data point x_i
- Choose table $j \propto N_j$ and Sample $x_i \sim f(\phi_j)$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1})$

The rich gets richer effect
CANNOT handle sequential data

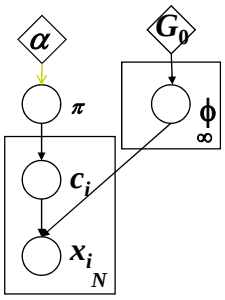
More formally …

- This is a Dirichlet Process Mixture Model
- Three equivalent constructions:



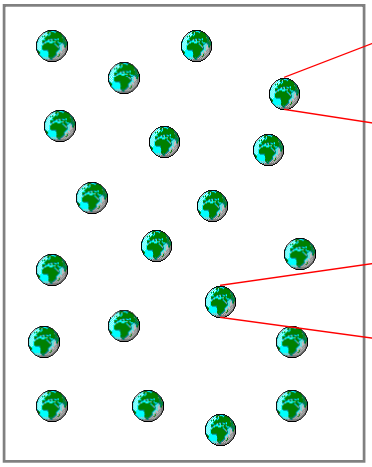
Measure over Measures

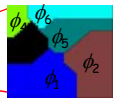
Chinese Restaurant Process



The Stick-breaking construction

Dirichlet Process





a distribution

another distribution

- 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

Stick-breaking Process

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

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

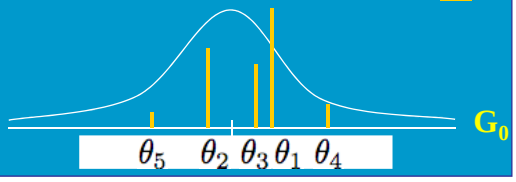
$$\theta_k \sim G_0$$ Location

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

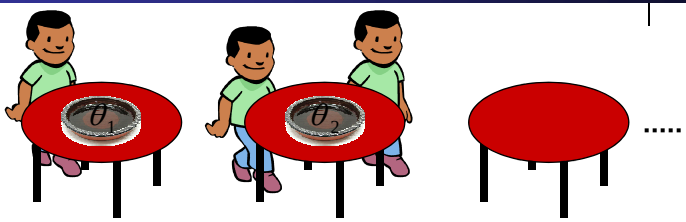
$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_k)$$

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

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



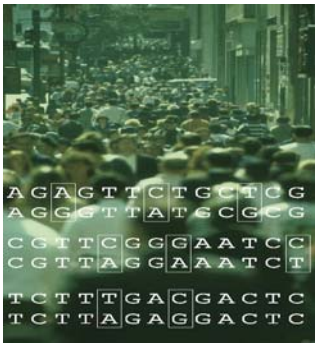
Chinese Restaurant Process

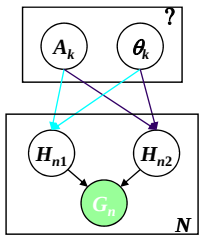


$$P(c_i = k \mid \mathbf{c}_{-i}) = \begin{array}{ccccc} 1 & & 0 & & 0 \\ \frac{1}{1+\alpha} & & \frac{\alpha}{1+\alpha} & & 0 \\ \frac{1}{2+\alpha} & & \frac{1}{2+\alpha} & & \frac{\alpha}{2+\alpha} \\ \frac{1}{3+\alpha} & & \frac{2}{3+\alpha} & & \frac{\alpha}{3+\alpha} \\ \frac{m_1}{i+\alpha-1} & & \frac{m_2}{i+\alpha-1} & \cdots & \frac{\alpha}{i+\alpha-1} \end{array}$$

CRP defines an exchangeable distribution on partitions over an (infinite) sequence of samples, such a distribution is formally known as the Dirichlet Process (DP)

Case study: ancestral Inference

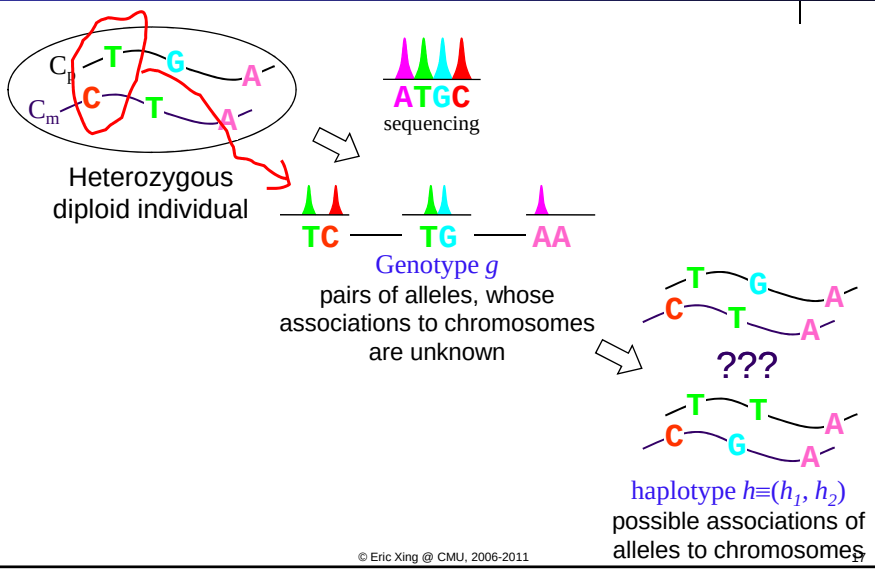




Essentially a clustering problem, but …

- Better recovery of the ancestors leads to better haplotyping results (because of more accurate grouping of common haplotypes)
- True haplotypes are obtainable with high cost, but they can validate model more subjectively (as opposed to examining saliency of clustering)
- Many other biological/scientific utilities

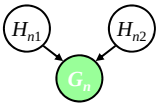
Background:haplotype ambiguity



A Finite (Mixture of) Allele Model

- The probability of a genotype g:

$$p(g) = \sum_{h_1,h_2 \in \mathcal{H}} p(h_1,h_2)\, p(g \mid h_1,h_2)$$

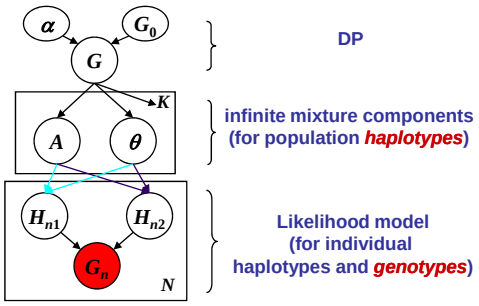


Population haplotype pool | Haplotype model | Genotyping model

- Standard settings:
  - $|\mathcal{H}| = K << 2^J$ fixed-sized population haplotype pool
  - $p(h_1,h_2) = p(h_1)p(h_2) = f_1 f_2$ Hardy-Weinberg equilibrium
- Problem: K ? $\mathcal{H}$?

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

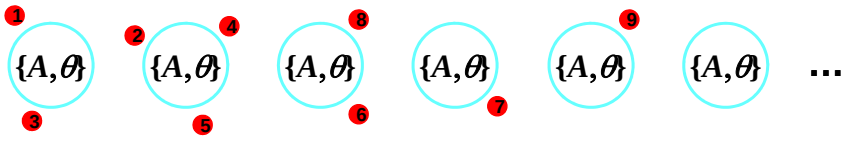
- Clustering human populations



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

The DP Mixture of Ancestral Haplotypes

- The customers around a table in CRP form a cluster
  - associate a mixture component (*i.e.*, a population haplotype) with a table
  - sample $\{a, \theta\}$ at each table from a base measure G_0 to obtain the population haplotype and nucleotide substitution frequency for that component



  - With $p(h|\{A, \theta\})$ and $p(g|h_1,h_2)$, the CRP yields a posterior distribution on the number of population haplotypes (and on the haplotype configurations and the nucleotide substitution frequencies)

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}

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

Posterior — Prior — x — Likelihood

CRP

$\vdots$

  - **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

$$\begin{aligned} & p(c_{i_e}^{(j)} = k | \mathbf{c}^{[-j,i_e]}, \mathbf{h}, \mathbf{a}) \\ & \propto p(c_{i_e}^{(j)} = k | \mathbf{c}^{[-j,i_e]}, \mathbf{m}, \mathbf{n}) p(h_{i_e}^{(j)} | a_k, \mathbf{c}, \mathbf{h}^{[-j,i_e]}) \\ & \propto (m_{jk}^{[-j,i_e]} + \tau\beta_k) p(h_{i_e}^{(j)} | a_k, \mathbf{l}_k^{[-j,i_e]}), \text{ for } k = 1, \ldots, K+1 \end{aligned}$$

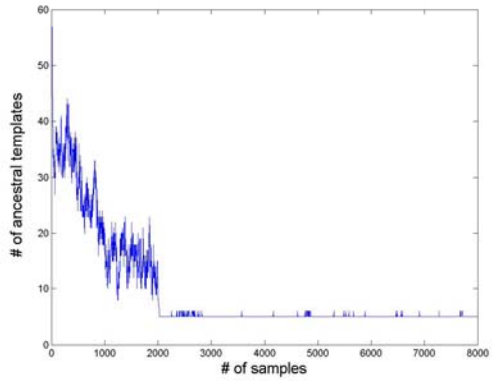
2. Sample a_k from

$$\begin{aligned} p(a_{k,t} | \mathbf{c}, \mathbf{h}) & \propto \prod_{j, i_e | c_{i_e,t}^{(j)} = k} p(h_{i_e,t}^{(j)} | a_{k,t}, l_{k,t}^{(j)}) \\ & = \frac{\Gamma(\alpha_h + l_{k,t}) \Gamma(\beta_h + l'_{k,t})}{\Gamma(\alpha_h + \beta_h + m_k)(|B|-1)^{l'_{k,t}}} R(\alpha_h, \beta_h) \end{aligned}$$

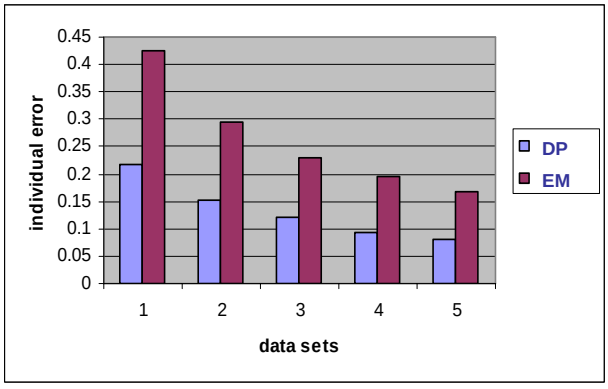
3. Sample $h_{ie}^{(j)}$ from $p(h_{i_e,t}^{(j)} | \mathbf{h}_{[-i_e,t]}^{(j)}, \mathbf{c}, \mathbf{a}, \mathbf{g})$

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

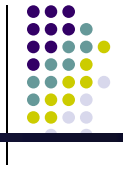
Convergence of Ancestral Inference



DP vs. Finite Mixture via EM



Outline



- Dirichlet Process: From finite mixture to Infinite mixture model
  - The Chinese Restaurant Process
  - Example: heliotype inference
- Intro to Markov Chain Monte Carlo
  - Gibbs sampling
- Dynamic Dirichlet Process
  - The recurrent Chinese Restaurant Process
  - Example: Application: evolutionary clustering of documents

Monte Carlo methods

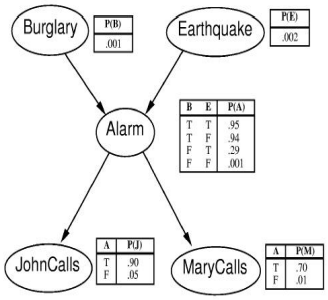
- Draw random samples from the desired distribution
- Yield a stochastic representation of a complex distribution
  - marginals and other expections can be approximated using sample-based averages

$$E[f(x)] = \frac{1}{N}\sum_{t=1}^{N} f(x^{(t)})$$

- **Asymptotically** exact and easy to apply to arbitrary models
- Challenges:
  - how to draw samples from a given dist. (not all distributions can be trivially sampled)?
  - how to make better use of the samples (not all sample are useful, or eqally useful, see an example later)?
  - how to know we've sampled enough?

Example: naive sampling

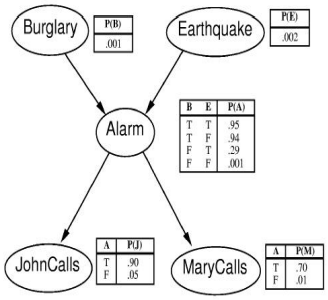
- Sampling: Construct samples according to probabilities given in a BN.



Alarm example: (Choose the right sampling sequence)
1) Sampling:P(B)=<0.001, 0.999> suppose it is false, B0. Same for E0. P(A|B0, E0)=<0.001, 0.999> suppose it is false...

Example: naive sampling

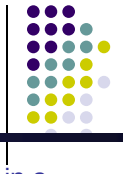
- Sampling: Construct samples according to probabilities given in a BN.



Alarm example: (Choose the right sampling sequence)
1) Sampling:P(B)=<0.001, 0.999> suppose it is false, B0. Same for E0. P(A|B0, E0)=<0.001, 0.999> suppose it is false...
2) Frequency counting: In the samples right, P(J|A0)=P(J,A0)/P(A0)=<1/9, 8/9>.

E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E1	B0	A1	M1	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0

Example: naive sampling



- Sampling: Construct samples according to probabilities given in a BN.

Alarm example: (Choose the right sampling sequence)

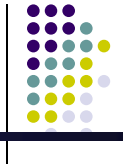
3) what if we want to compute P(J|A1) ?
we have only one sample ...
P(J|A1)=P(J,A1)/P(A1)=<0, 1>.

4) what if we want to compute P(J|B1) ?
No such sample available!
P(J|A1)=P(J,B1)/P(B1) can not be defined.

For a model with hundreds or more variables, rare events will be very hard to garner evough samples even after a long time or sampling ...

E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E1	B0	A1	M1	J1
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0
E0	B0	A0	M0	J0

Monte Carlo methods (cond.)



- Direct Sampling
  - We have seen it.
  - Very difficult to populate a high-dimensional state space
- Rejection Sampling
  - Create samples like direct sampling, only count samples which is consistent with given evidences.
- ….
- Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo

- Samples are obtained from a **Markov chain** (of sequentially evolving distributions) whose **stationary distribution** is the desired $p(x)$
- Construct a Markov chain whose stationary distribution is the target density $= P(X|e)$.
- Run for T samples (burn-in time) until the chain converges/mixes/reaches stationary distribution.
- Then collect M (correlated) samples x_m .
- Key issues:
  - Designing proposals so that the chain mixes rapidly.
  - Diagnosing convergence.

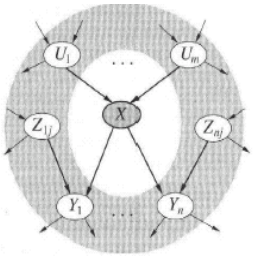
Gibbs sampling

- Gibbs sampling is an MCMC algorithm that is especially appropriate for inference in graphical models.
- The procedue
  - we have variable set $\boldsymbol{X}=\{x_1, x_2, x_3, \ldots x_N\}$ for a GM
  - at each step one of the variables X_i is selected (at random or according to some fixed sequences), denote the remaining variables as $\boldsymbol{X}_{-i}$, and its current value as $\boldsymbol{x}_{-i}^{(t-1)}$
    - Using the "alarm network" as an example, say at time t we choose X_E, and we denote the current value assignments of the remaining variables, $\boldsymbol{X}_{-E}$, obtained from previous samples, as $\boldsymbol{x}_{-E}^{(t-1)} = \{x_B^{(t-1)}, x_A^{(t-1)}, x_J^{(t-1)}, x_M^{(t-1)}\}$
  - the conditonal distribution $p(X_i | \boldsymbol{x}_{-i}^{(t-1)})$ is computed
  - a value $\boldsymbol{x}_i^{(t)}$ is sampled from this distribution
  - the sample $\boldsymbol{x}_i^{(t)}$ replaces the previous sampled value of X_i in $\boldsymbol{X}$.
    - i.e., $\boldsymbol{x}^{(t)} = \boldsymbol{x}_{-E}^{(t-1)} \cup \boldsymbol{x}_E^{(t)}$

Why Gibbs Sampling is Simple

- Markov-Blanket
  - A variable is independent from others, given its parents, children and children's parents. d-separation.

    $\Rightarrow p(X_i|\boldsymbol{X}_{-i}) = p(X_i|\text{MB}(X_i))$

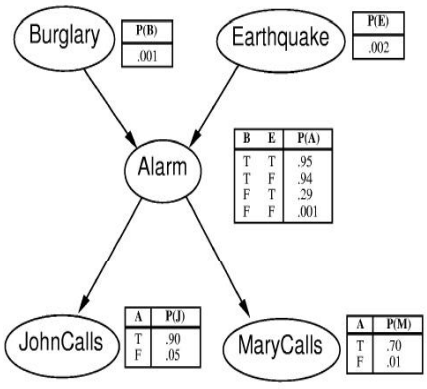


- Gibbs sampling
  - Create a random sample. Every step, choose one variable and sample it by P(X|MB(X)) based on previous sample.

MB(A)={B, E, J, M}

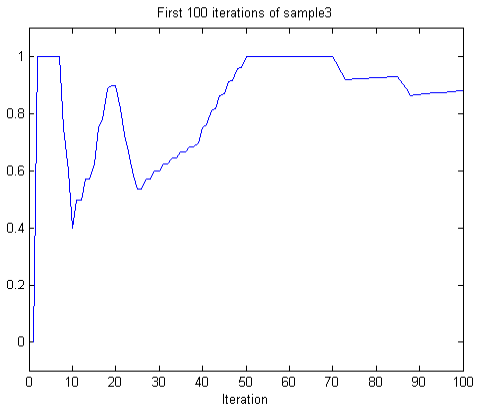
MB(E)={A, B}

Example: alarm network again

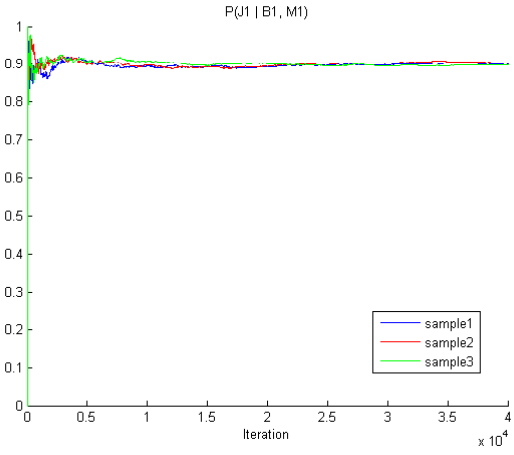


- To calculate P(J|B1,M1)
- Choose (B1,E0,A1,M1,J1) as a start
- **Evidences** are B1, M1, **variables** are A, E, J.
- Choose next variable as A
- Sample A by P(A|MB(A))=P(A|B1, E0, M1, J1) suppose to be false.
- (B1, E0, A0, M1, J1)
- Choose next random variable as E, sample E~P(E|B1,A0)
- ...

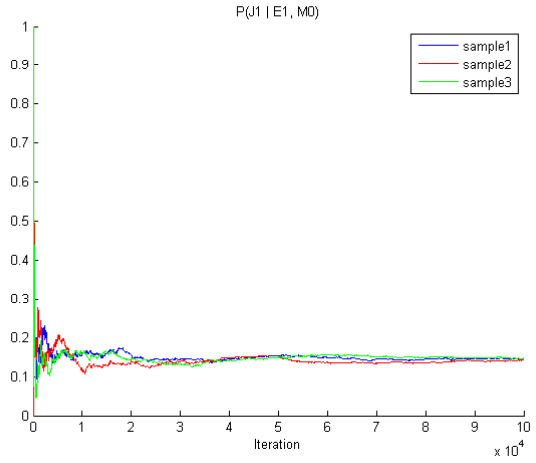
Example



Example:

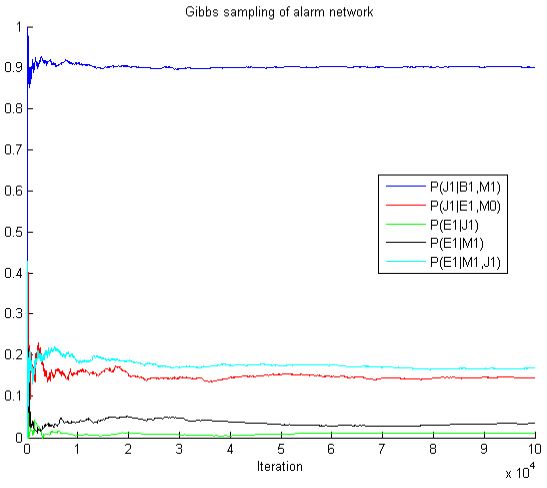


Example



Example

P(J1 | B1,M1) = 0.90
P(J1 | E1,M0) = 0.14
P(E1 | J1) = 0.01
P(E1 | M1) = 0.04
P(E1 | M1,J1) = 0.17

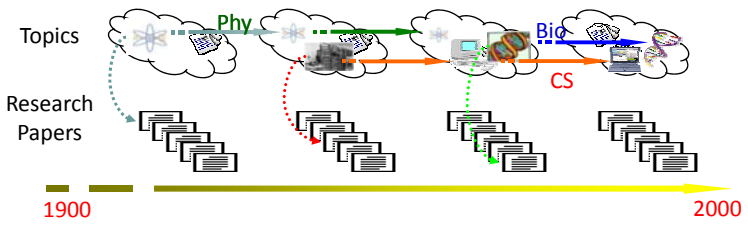


Outline

- Dirichlet Process: From finite mixture to Infinite mixture model
  - The Chinese Restaurant Process
  - Example: heliotype inference
- Intro to Markov Chain Monte Carlo
  - Gibbs sampling
- Dynamic Dirichlet Process
  - The recurrent Chinese Restaurant Process
  - Example: Application: evolutionary clustering of documents

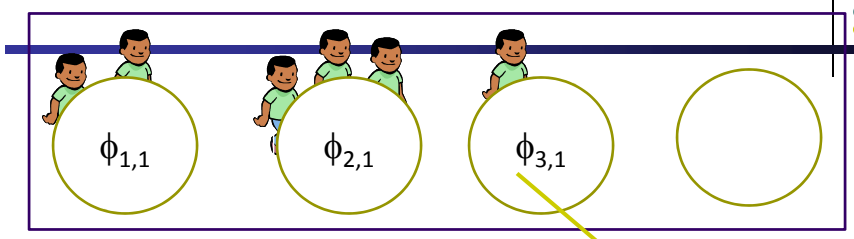
Evolutionary Clustering

- 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



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.

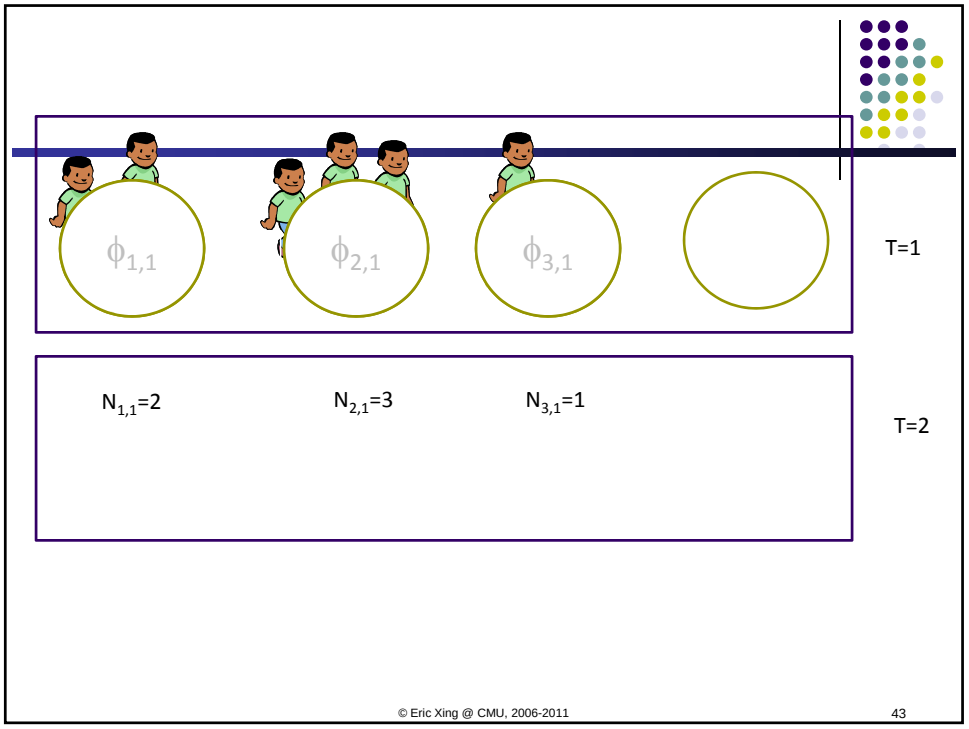


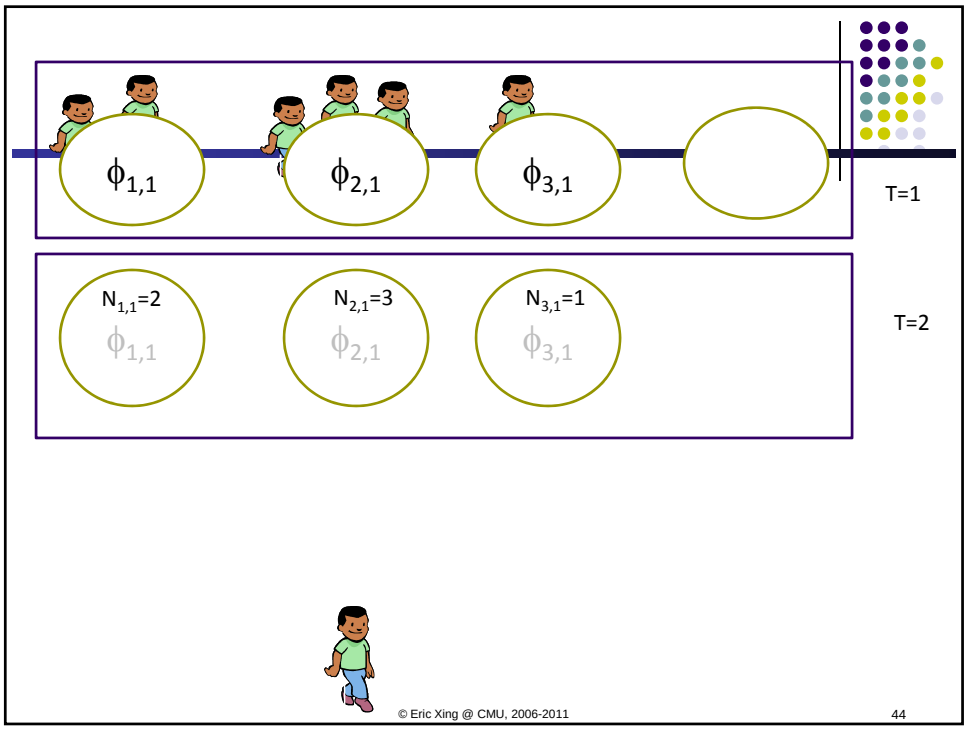
T=1

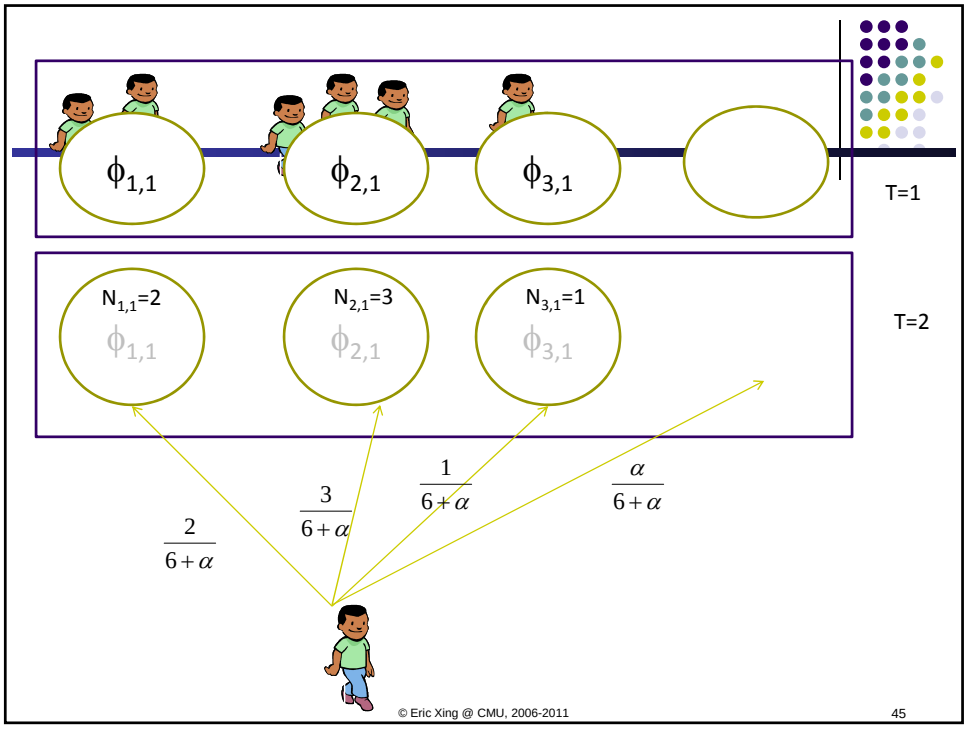
Dish eaten at table 3 at time epoch 1
OR the parameters of cluster 3 at time epoch 1

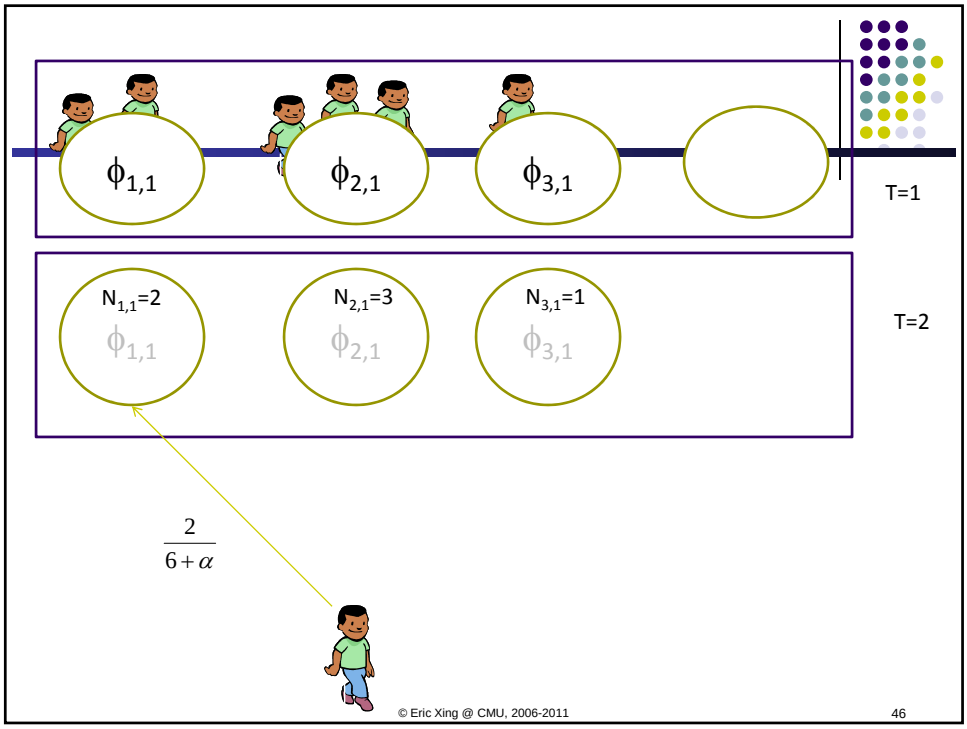
Generative Process

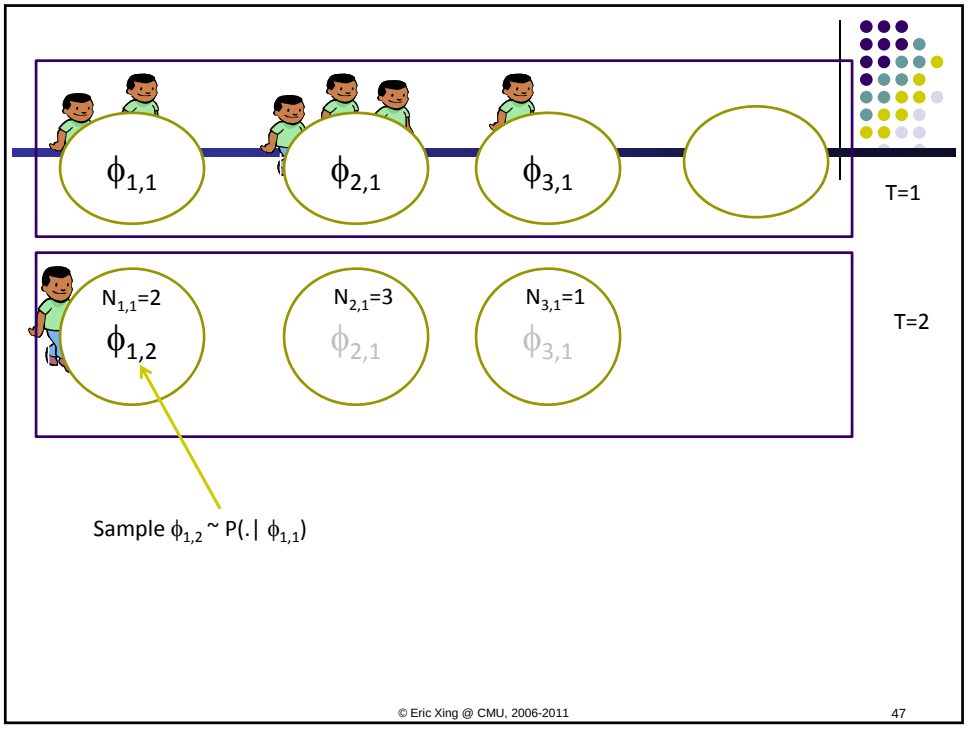
-Customers at time T=1 are seated as before:
- Choose table $j \propto N_{j,1}$ and Sample $x_i \sim f(\phi_{j,1})$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1,1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1,1})$

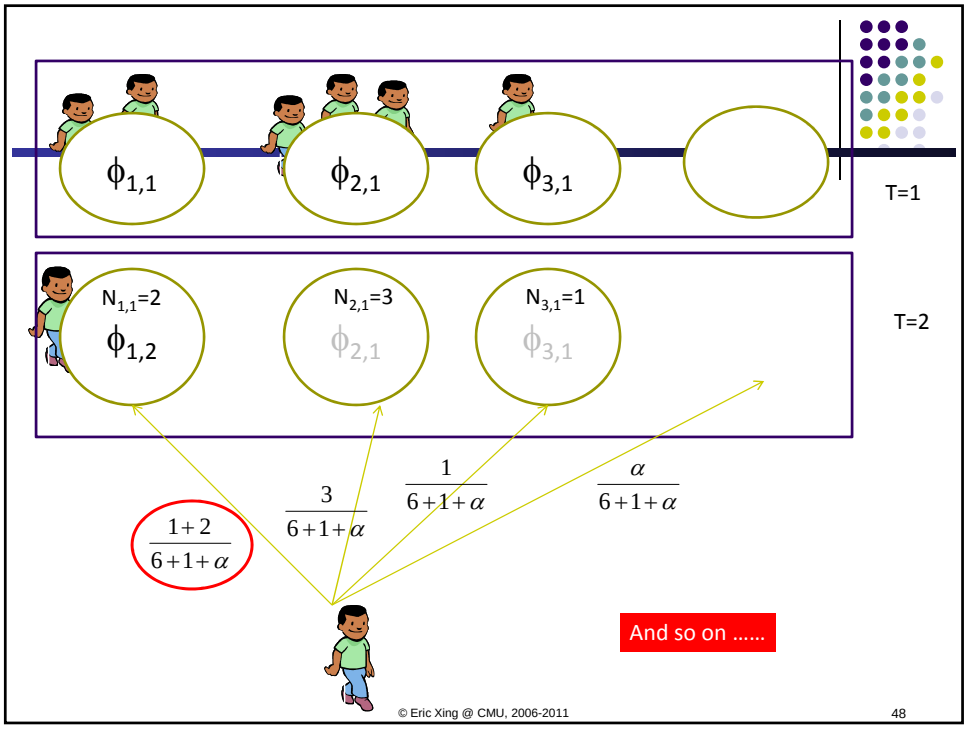

$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$N_{1,1}$=2
$N_{2,1}$=3
$N_{3,1}$=1
T=2

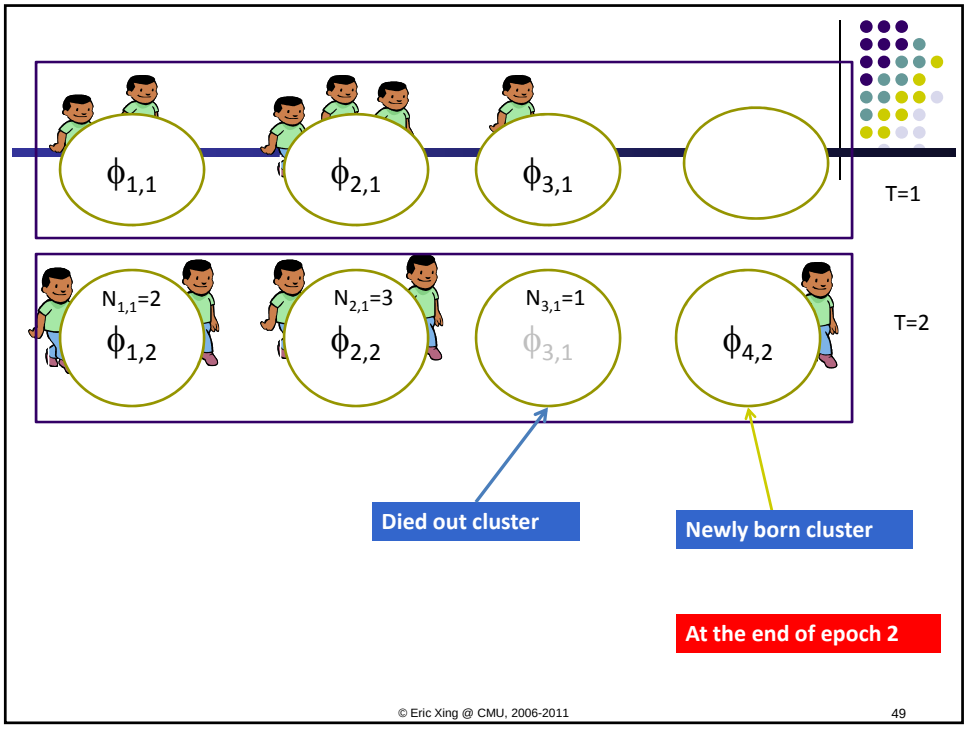

$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$N_{1,1}$=2
$N_{2,1}$=3
$N_{3,1}$=1
T=2

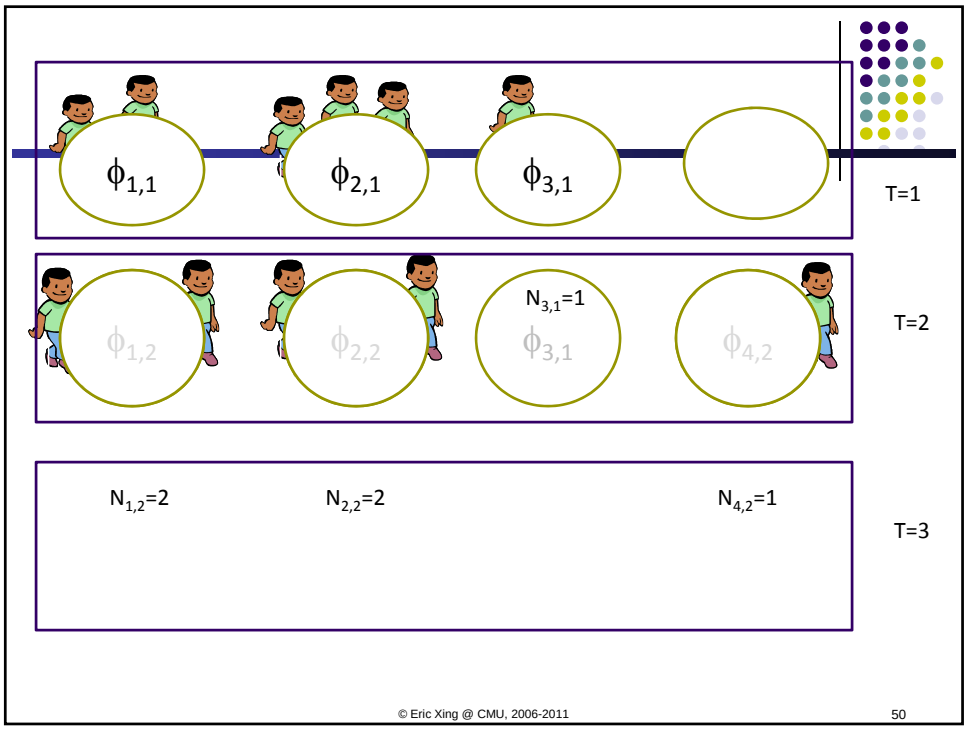












Temporal DPM

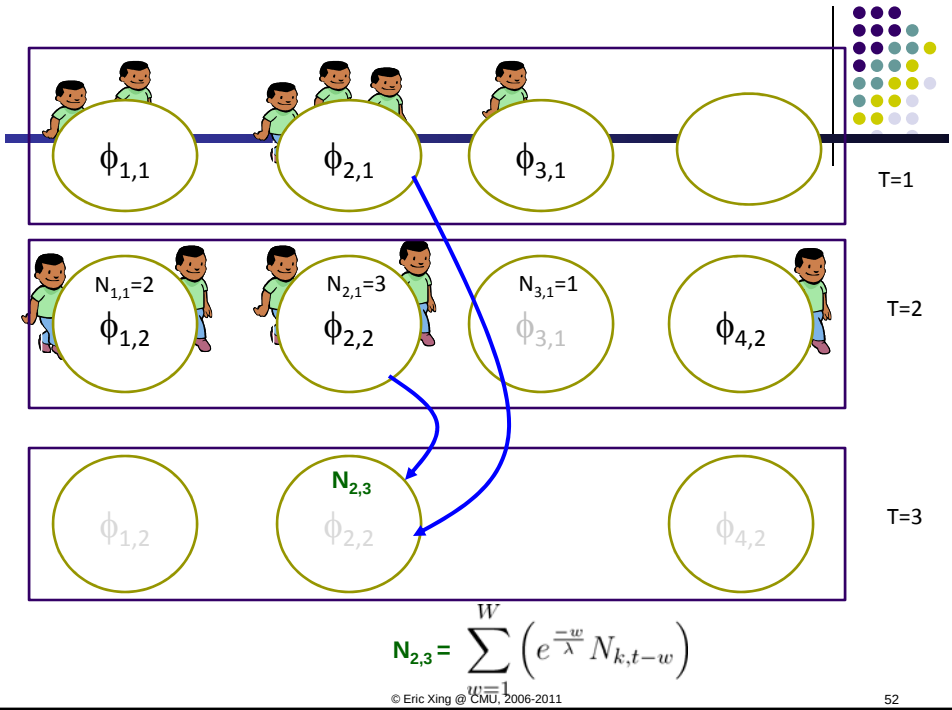
- Can be extended to model higher-order dependencies
- Can decay dependencies over time
  - Pseudo-counts for table *k* at time *t is*

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

History size

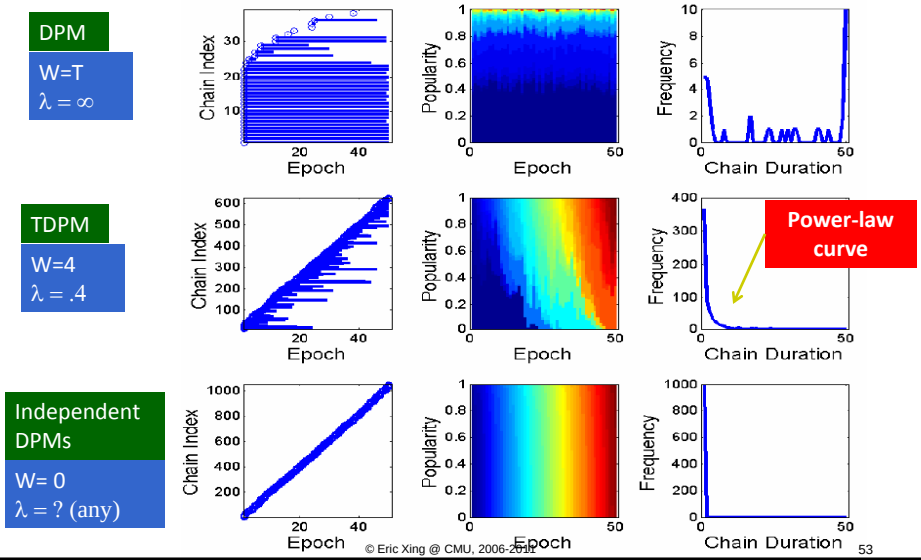
Decay factory

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



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

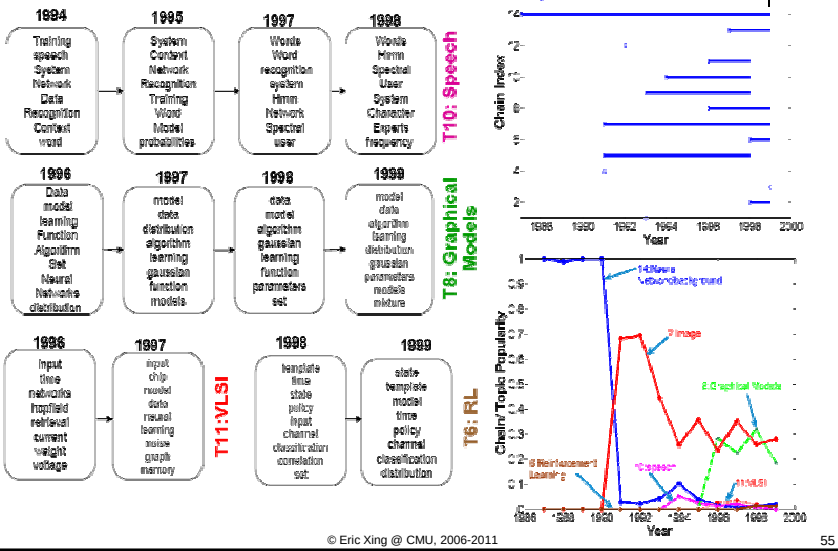
TDPM Generative Power



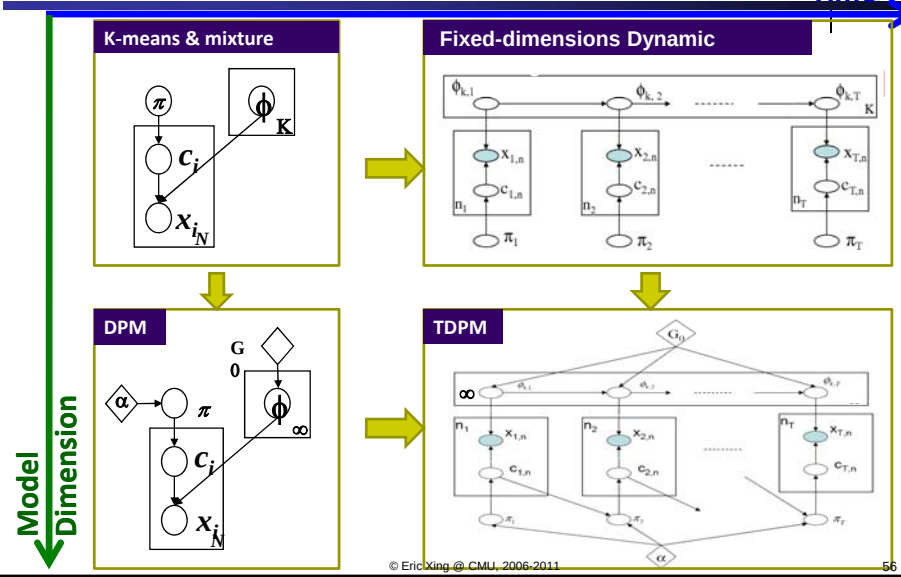
Results: NIPS 12

- 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 NIPS trends



The Big Picture



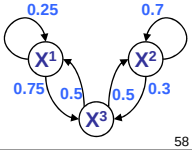
Appendix:

Theory of MCMC (optional)

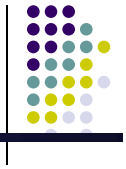
- **Definition:** Markov Chains
  - Given an n-dimensional state space
  - Random vector $\mathbf{X} = (x_1,\ldots,x_n)$
  - $\mathbf{x}^{(t)} = \mathbf{x}$ at time-step t
  - $\mathbf{x}^{(t)}$ transitions to $\mathbf{x}^{(t+1)}$ with prob
    $P(\mathbf{x}^{(t+1)} \mid \mathbf{x}^{(t)},\ldots,\mathbf{x}^{(1)}) = T(\mathbf{x}^{(t+1)} \mid \mathbf{x}^{(t)}) = T(\mathbf{x}^{(t)} \rightarrow \mathbf{x}^{(t+1)})$
- **Homogenous**: chain determined by state $\mathbf{x}^{(0)}$, fixed *transition kernel* T (rows sum to 1)
- **Equilibrium**: $\pi(\mathbf{x})$ is a *stationary (equilibrium) distribution* if
  $\pi(\mathbf{x}') = \Sigma_{\mathbf{x}} \pi(\mathbf{x})\ T(\mathbf{x}\rightarrow\mathbf{x}')$.

  i.e., is a left eigenvector of the transition matrix $\pi^T T = \pi^T$.

$$(0.2 \quad 0.5 \quad 0.3) = (0.2 \quad 0.5 \quad 0.3)\begin{pmatrix} 0.25 & 0 & 0.75 \\ 0 & 0.7 & 0.3 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$



Markov Chains



- An MC is ***irreducible*** if transition graph connected
- An MC is ***aperiodic*** if it is not trapped in cycles
- An MC is ***ergodic*** (regular) if you can get from state x to x' in a finite number of steps.
- **Detailed balance**: prob($\mathbf{x}^{(t)} \rightarrow \mathbf{x}^{(i-1)}$) = prob($\mathbf{x}^{(t-1)} \rightarrow \mathbf{x}^{(t)}$)

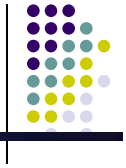
$$p(\mathbf{x}^{(t)})\mathcal{T}(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)}) = p(\mathbf{x}^{(t-1)})\mathcal{T}(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)})$$

summing over $\mathbf{x}^{(t-1)}$

$$p(\mathbf{x}^{(t)}) = \sum_{\mathbf{x}^{(t-1)}} p(\mathbf{x}^{(t-1)})\mathcal{T}(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)})$$

- Detailed bal → stationary dist exists

MCMC Via Metropolis-Hastings



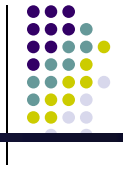
- Treat the target distribution as stationary distribution
- Sample from an easier proposal distribution, followed by an acceptance test
- This induces a transition matrix that satisfies detailed balance
  - MH proposes moves according to $Q(x'|x)$ and accepts samples with probability $A(x'|x)$.
  - The induced transition matrix is $\mathcal{T}(x \rightarrow x') = Q(x'|x)A(x'|x)$
  - Detailed balance means

$$\pi(x)Q(x'|x)A(x'|x) = \pi(x')Q(x|x')A(x|x')$$

  - Hence the acceptance ratio is

$$A(x'|x) = \min\left(1, \frac{\pi(x')Q(x|x')}{\pi(x)Q(x'|x)}\right)$$

Gibbs sampling



- Gibbs sampling is a special case of MH
- The transition matrix updates each node one at a time using the following proposal:

$$Q\left((x_i, \mathbf{x}_{-i}) \to (x_i', \mathbf{x}_{-i})\right) = p(x_i' \mid \mathbf{x}_{-i})$$

- This is efficient since for two reasons
  - It leads to samples that is always accepted

$$A\left((x_i, \mathbf{x}_{-i}) \to (x_i', \mathbf{x}_{-i})\right) = \min\left(1, \frac{p(x_i', \mathbf{x}_{-i}) Q\left((x_i', \mathbf{x}_{-i}) \to (x_i, \mathbf{x}_{-i})\right)}{p(x_i, \mathbf{x}_{-i}) Q\left((x_i, \mathbf{x}_{-i}) \to (x_i', \mathbf{x}_{-i})\right)}\right)$$

$$= \min\left(1, \frac{p(x_i' \mid \mathbf{x}_{-i}) p(\mathbf{x}_{-i}) p(x_i \mid \mathbf{x}_{-i})}{p(x_i \mid \mathbf{x}_{-i}) p(\mathbf{x}_{-i}) p(x_i' \mid \mathbf{x}_{-i})}\right) = \min(1,1)$$

  Thus $$T\left((x_i, \mathbf{x}_{-i}) \to (x_i', \mathbf{x}_{-i})\right) = p(x_i' \mid \mathbf{x}_{-i})$$

  - It is efficient since $p(x_i' \mid \mathbf{x}_{-i})$ only depends on the values in X_i's Markov blanket