Introduction to Bayesian Nonparametrics

Yee Whye Teh
Gatsby Computational Neuroscience Unit, UCL

MLSS 2011 Bordeaux
September 2011

Bayesian Machine Learning

Probabilistic Machine Learning

- Machine Learning is all about data.
  - Stochastic, chaotic and/or complex process
  - Noisily observed
  - Partially observed
- **Probability theory** is a rich language to express these uncertainties.
  - **Probabilistic models**
- Graphical tool to visualize complex models for complex problems.
- Complex models can be built from simpler parts.
- Computational tools to derive algorithmic solutions.
- Separation of modelling questions from algorithmic questions.

Probabilistic Modelling

- Data: $x_1, x_2, \ldots, x_n$.
- Latent variables: $y_1, y_2, \ldots, y_n$.
- Parameter: θ.
- A probabilistic model is a parametrized joint distribution over variables.

$$P(x_1, \ldots, x_n, y_1, \ldots, y_n | \theta)$$

- Typically interpreted as a **generative model** of data.
- Inference, of latent variables given observed data:

$$P(y_1, \ldots, y_n | x_1, \ldots, x_n, \theta) = \frac{P(x_1, \ldots, x_n, y_1, \ldots, y_n | \theta)}{P(x_1, \ldots, x_n | \theta)}$$

Probabilistic Modelling

- Learning, typically by maximum likelihood:

$$\theta^{\mathrm{ML}} = \operatorname*{argmax}_{\theta} P(x_1, \ldots, x_n | \theta)$$

- Prediction:

$$P(x_{n+1}, y_{n+1} | x_1, \ldots, x_n, \theta)$$

- Classification:

$$\operatorname*{argmax}_{c} P(x_{n+1} | \theta^c)$$

- Visualization, interpretation, summarization.
- Standard algorithms: EM, junction tree, variational inference, MCMC...

Bayesian Modelling

- Prior distribution:

$$P(\theta)$$

- Posterior distribution (both inference and learning):

$$P(y_1, \ldots, y_n, \theta | x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n, y_1, \ldots, y_n | \theta) P(\theta)}{P(x_1, \ldots, x_n)}$$

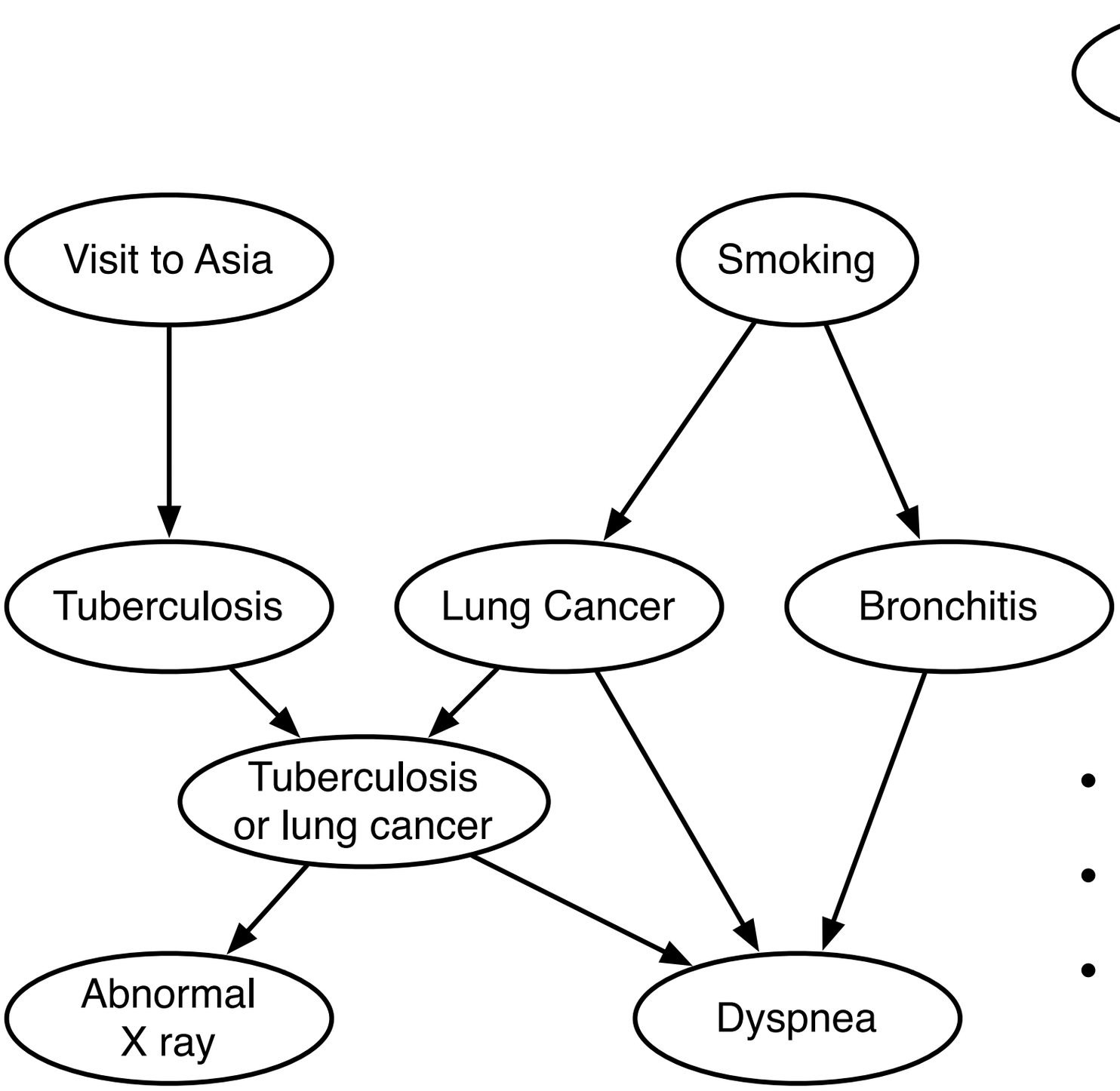
- Prediction:

$$P(x_{n+1} | x_1, \ldots, x_n) = \int P(x_{n+1} | \theta) P(\theta | x_1, \ldots, x_n) d\theta$$

- Classification:

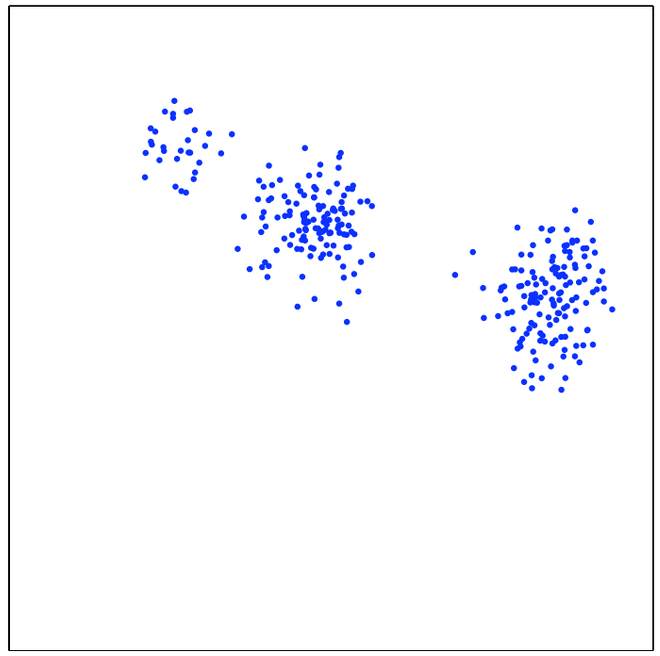
$$P(x_{n+1} | x_1^c, \ldots, x_n^c) = \int P(x_{n+1} | \theta^c) P(\theta^c | x_1^c, \ldots, x_n^c) d\theta^c$$

Graphical Models



- Nodes = variables
- Edges = dependencies
- Lack of edges = conditional independencies

Model-based Clustering



- Model for data from heterogeneous unknown sources.
- Each cluster (source) modelled using a parametric model (e.g. Gaussian).
- Data item *i*:

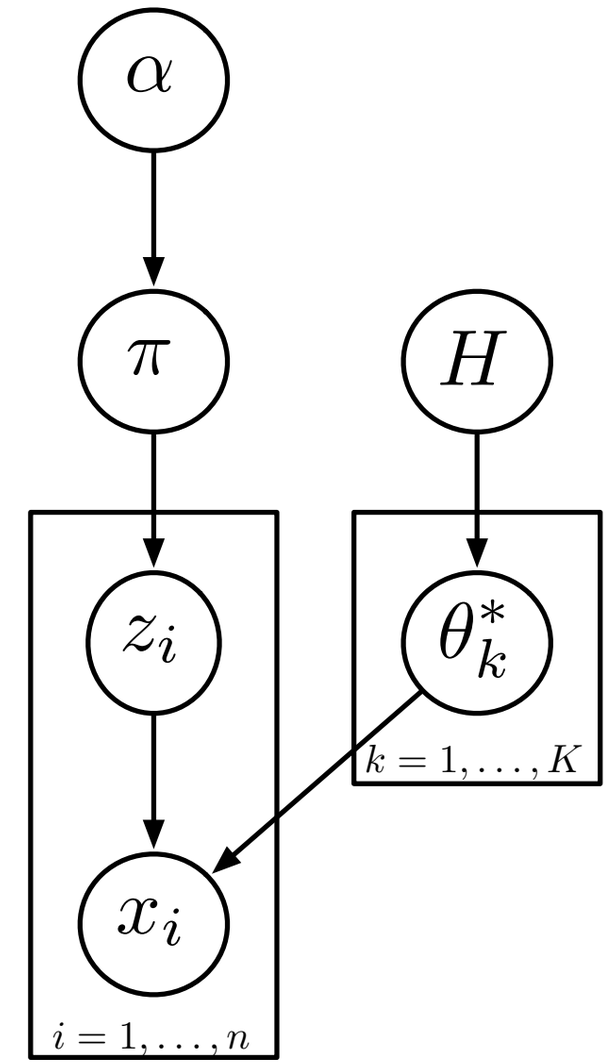
$$z_i|\pi \sim \text{Discrete}(\pi)$$
$$x_i|z_i, \theta^*_k \sim F(\theta^*_{z_i})$$

- Mixing proportions:

$$\pi = (\pi_1, \ldots, \pi_K)|\alpha \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

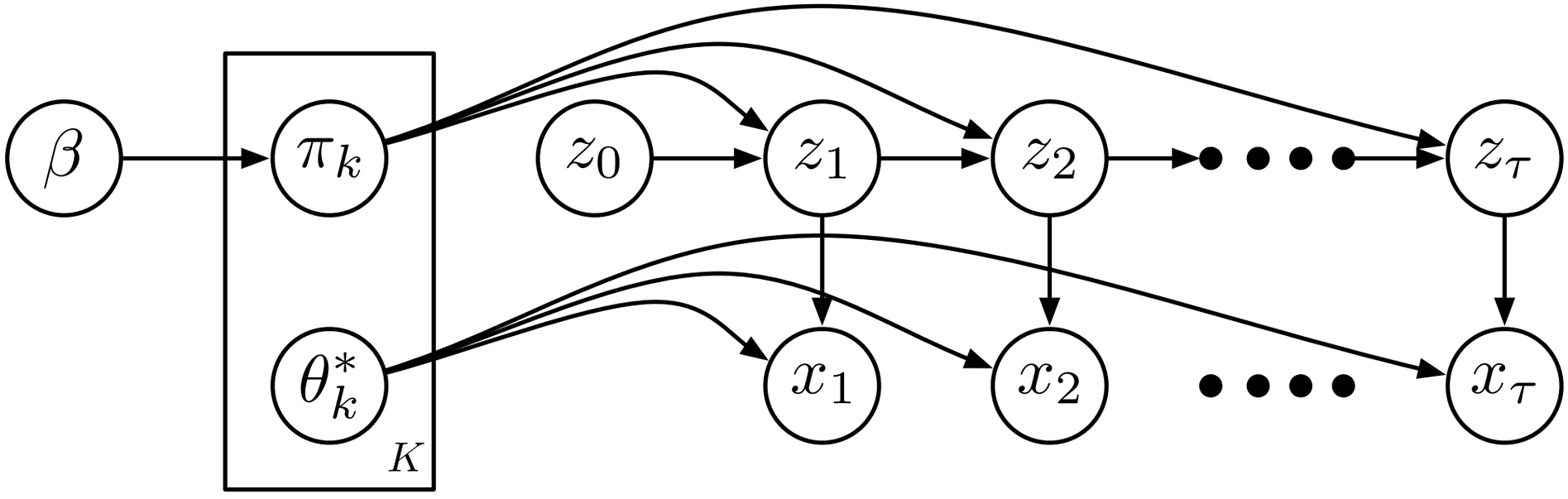
- Cluster *k*:

$$\theta^*_k|H \sim H$$



Hidden Markov Models

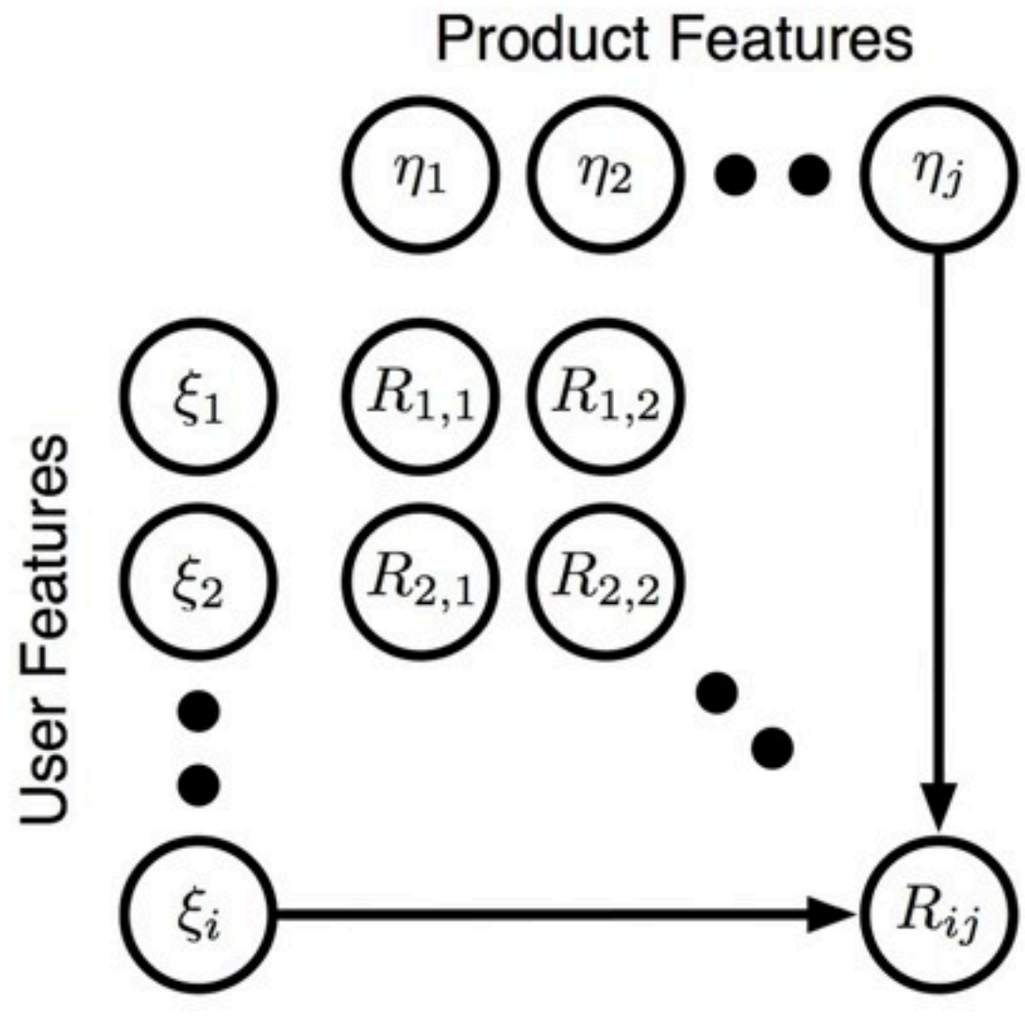
- Popular model for time series data.
- Unobserved dynamics modelled using a Markov model
- Observations modelled as independent conditioned on current state.



Collaborative Filtering

- Data: for each user i ratings R_{ij} for a subset of products j.
- Problem: predict how much users would like products that they haven't seen.

$$R_{ij}|\xi_i, \eta_j \sim \mathcal{N}(\xi_i^\top \eta_j, \sigma^2)$$



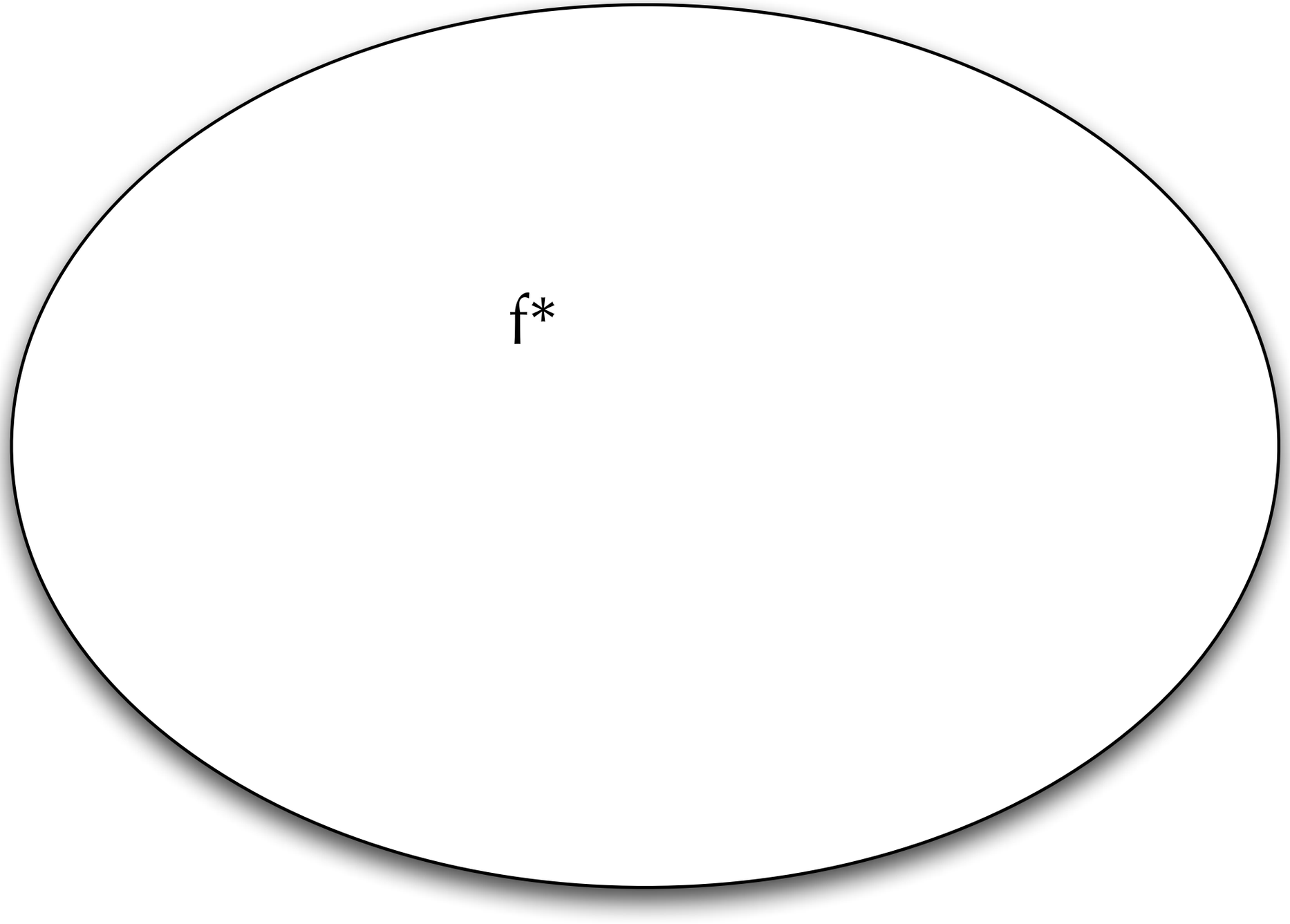
Bayesian Nonparametrics

[Hjort et al 2010]

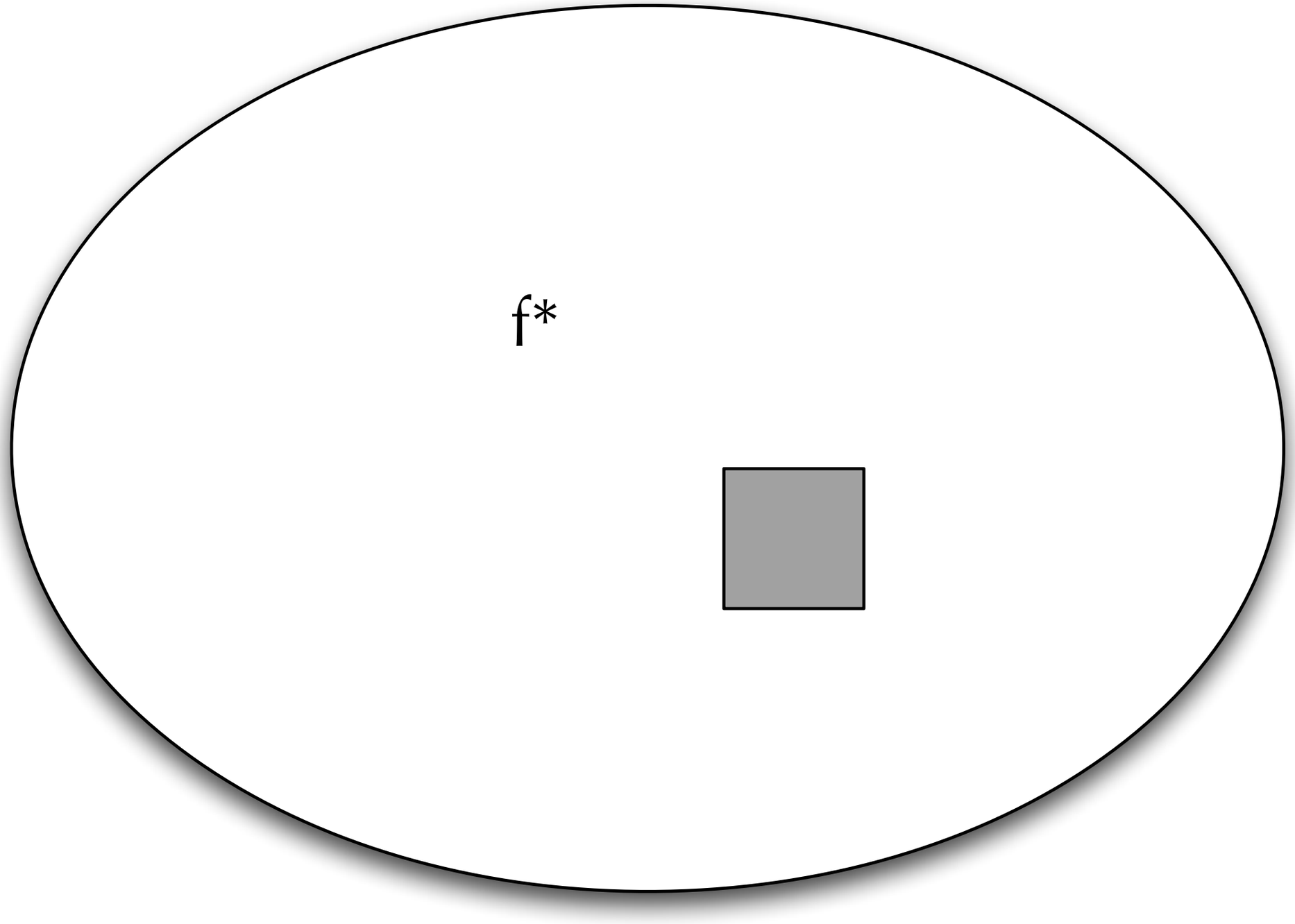
Bayesian Nonparametrics

- What is a nonparametric model?
  - A really large parametric model;
  - A parametric model where the number of parameters increases with data;
  - A family of distributions that is dense in some large space relevant to the problem at hand.

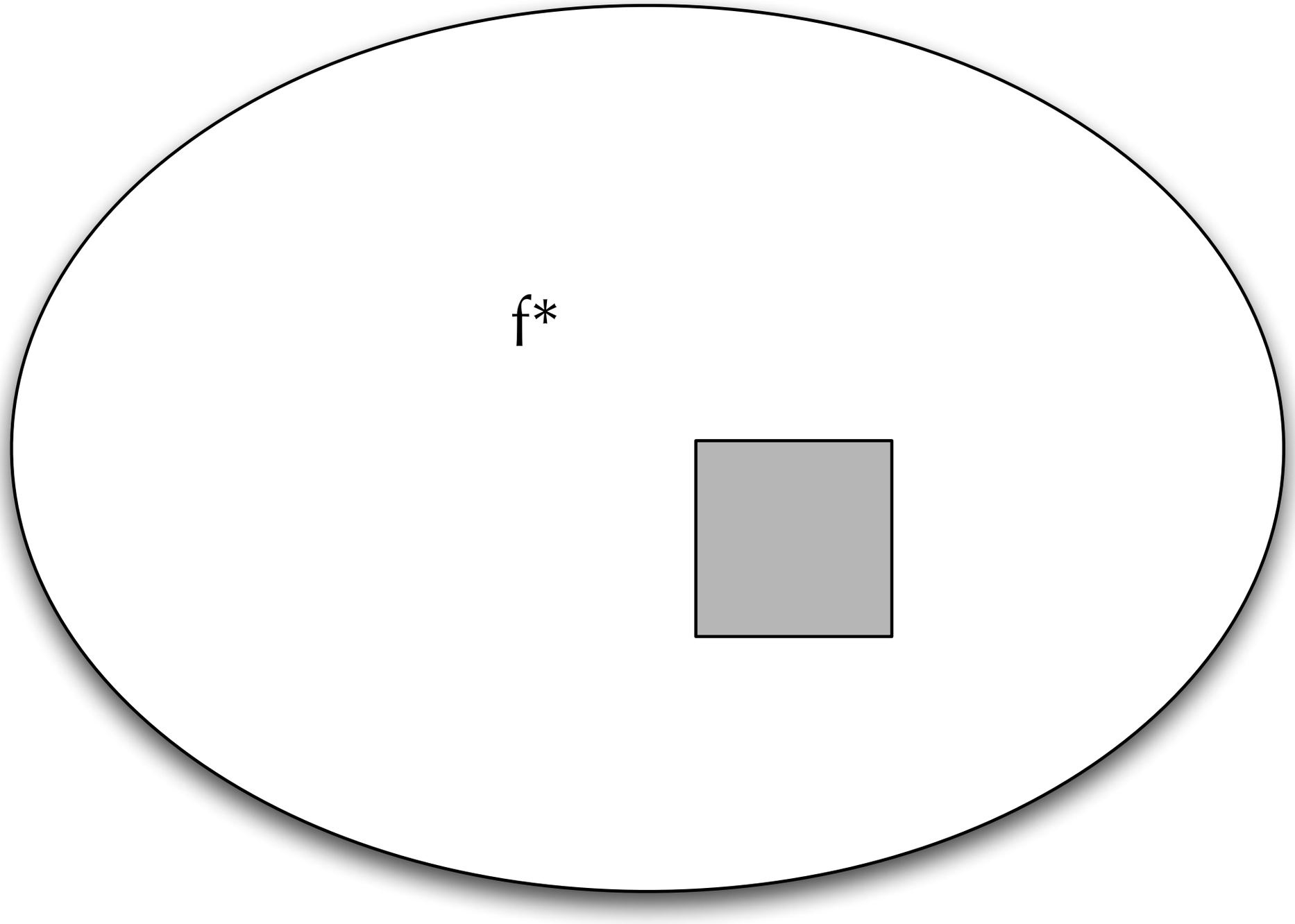
Bayesian Nonparametrics



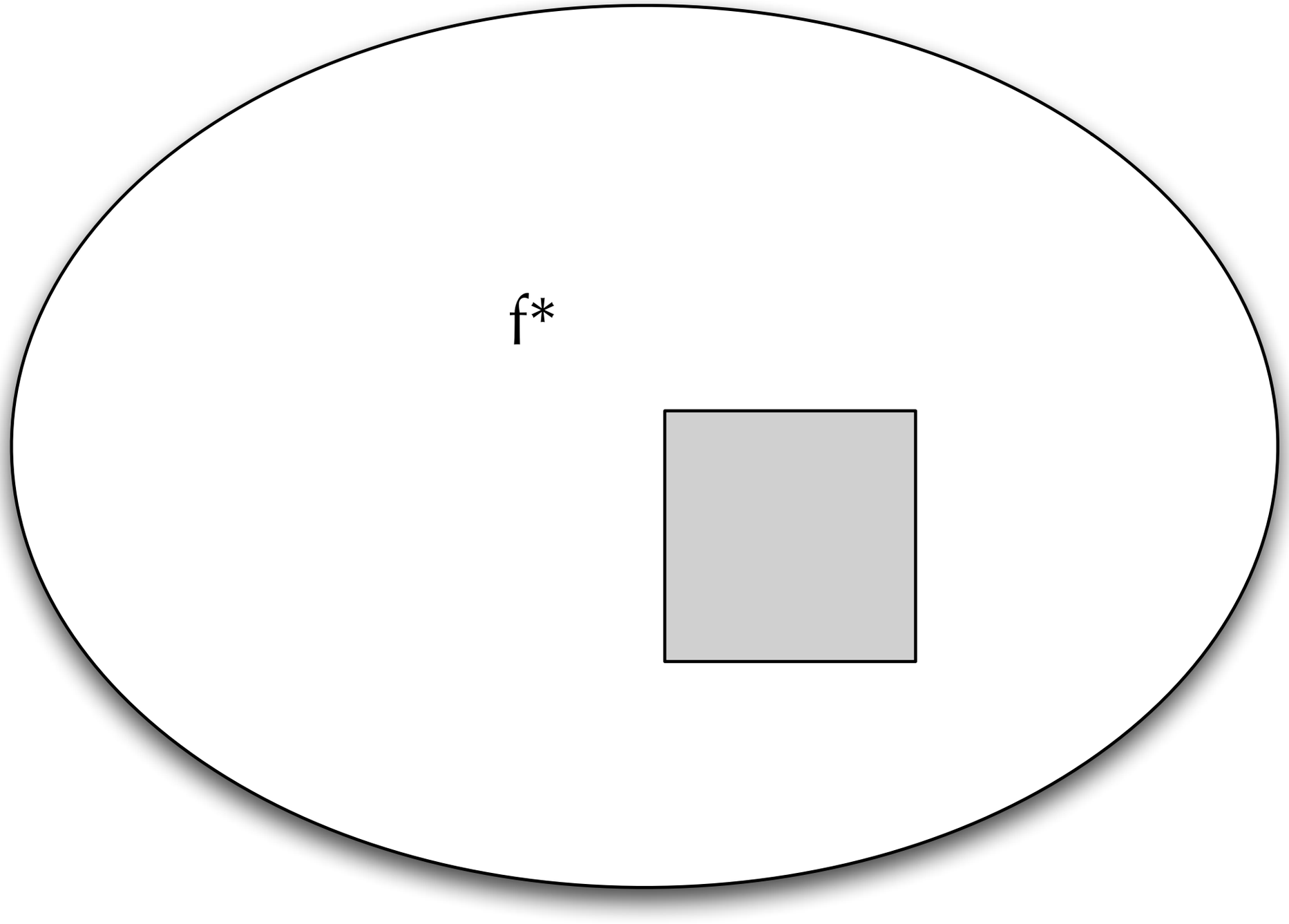
Bayesian Nonparametrics



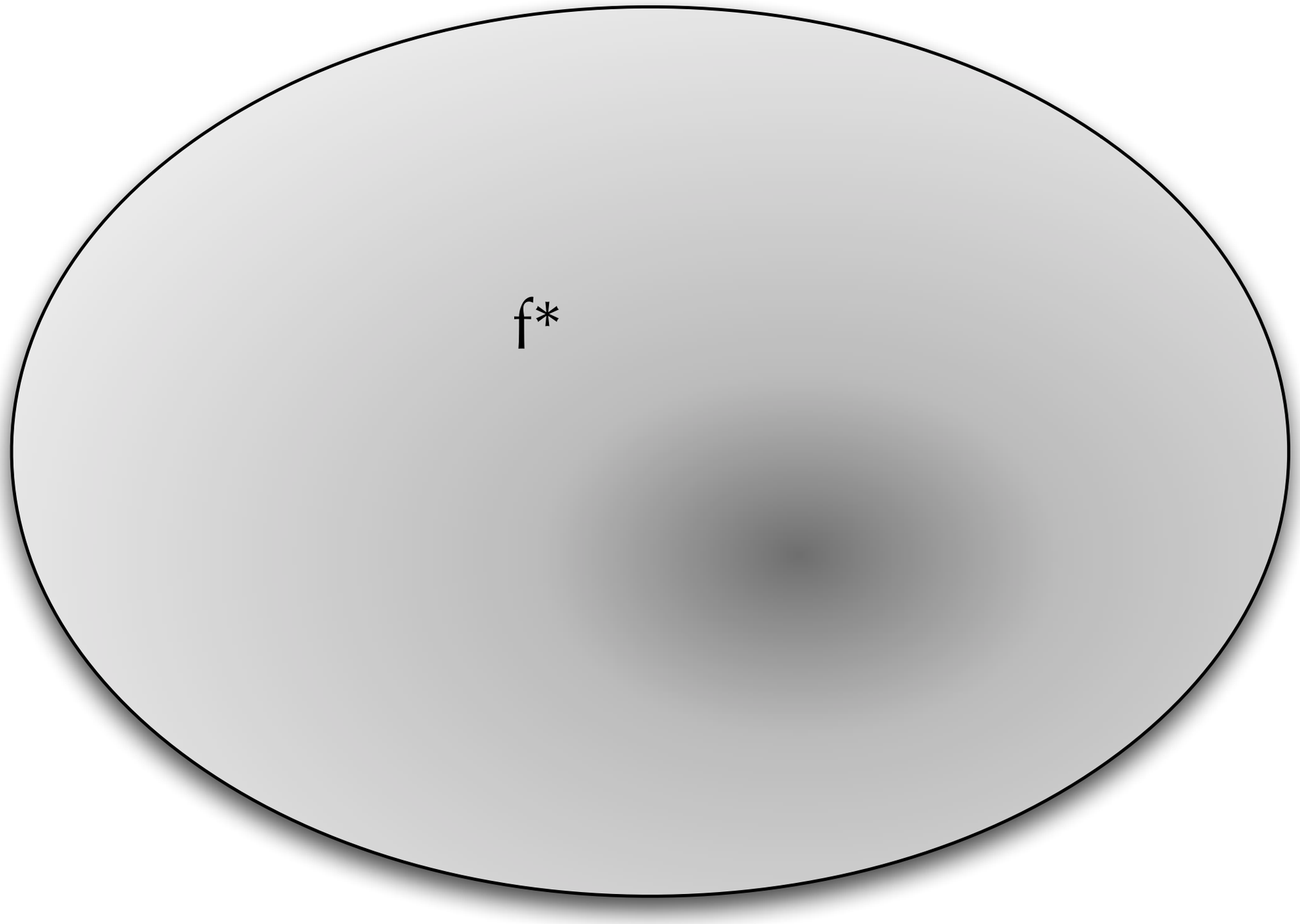
Bayesian Nonparametrics



Bayesian Nonparametrics



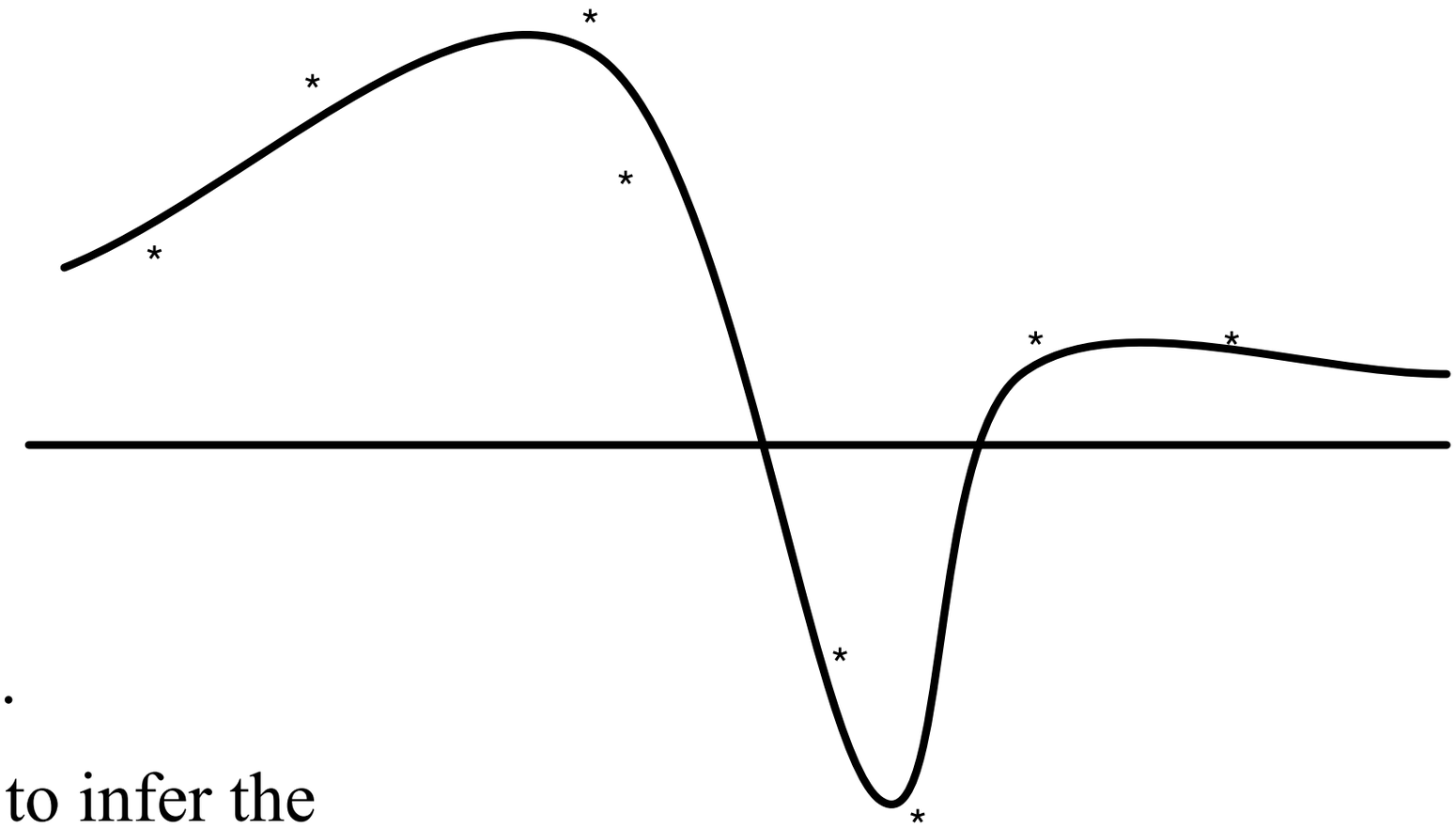
Bayesian Nonparametrics



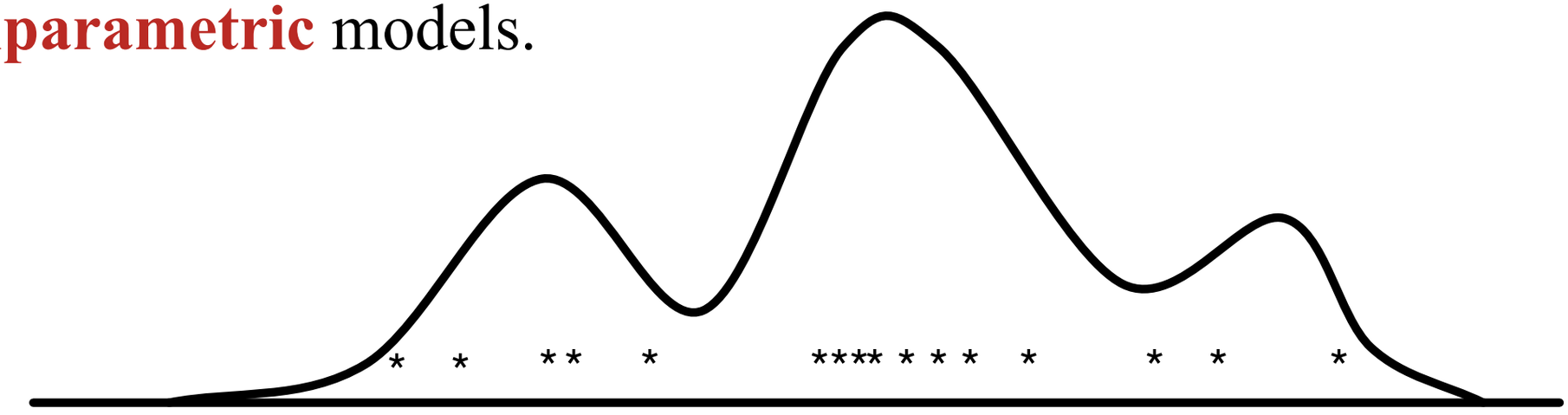
Reason 1: Model Selection and Averaging

- Model selection/averaging typically very expensive computationally.
- Used to prevent overfitting and underfitting.
- But a well-specified Bayesian model should not overfit anyway.
- By using a very large Bayesian model or one that grows with amount of data, we will not underfit either.
  - **Bayesian nonparametric** models.

Reason 2: Large Function Spaces

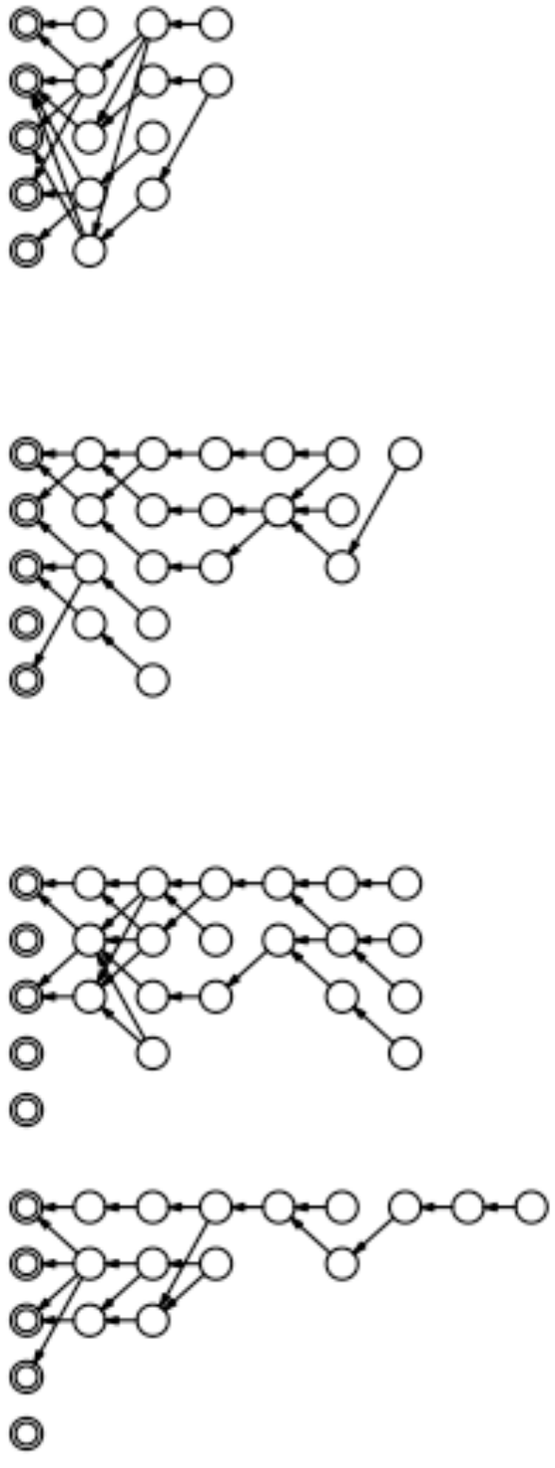


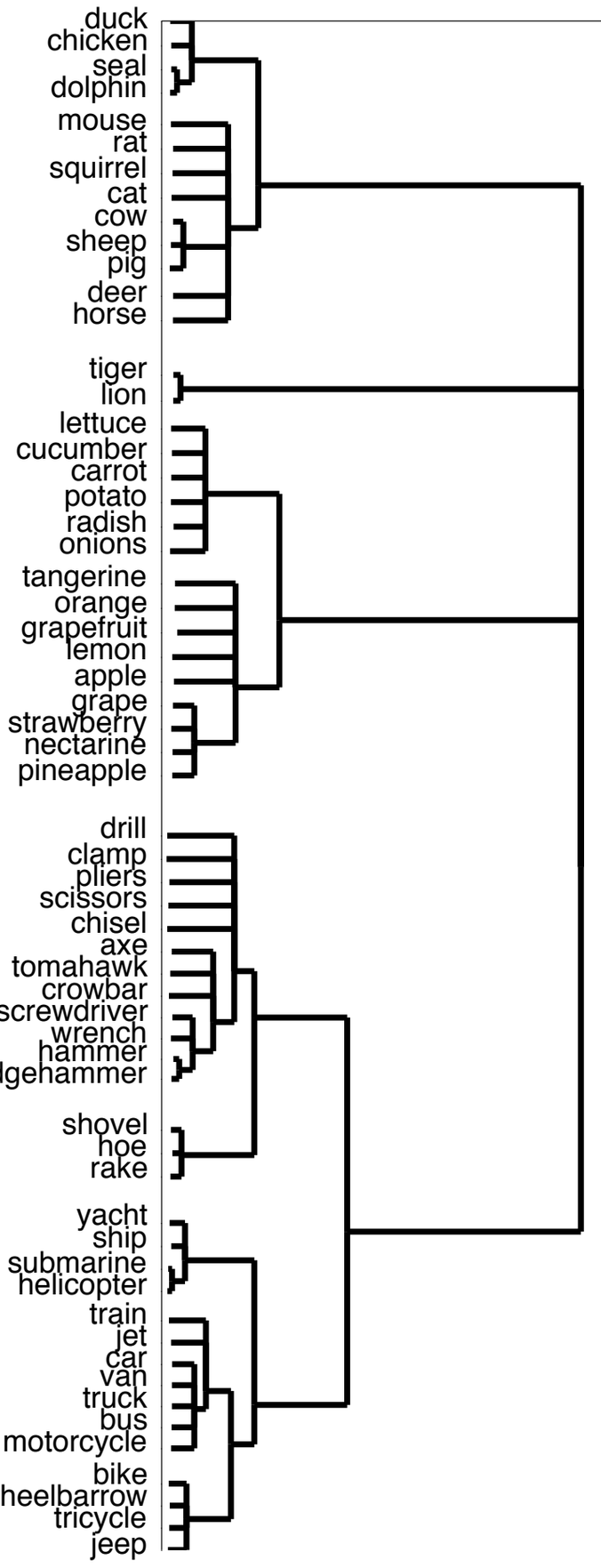
- Large function spaces.
- More straightforward to infer the infinite-dimensional objects themselves.
  - **Bayesian nonparametric** models.



Reason 3: Structural Learning

- Learning structures.
- Bayesian prior over combinatorial structures.
- Nonparametric priors sometimes end up simpler than parametric priors.





[Adams et al 2010, Blundell et al 2010]

Reason 4: Novel and Useful Properties

- Many interesting Bayesian nonparametric models with interesting and useful properties:
  - Projectivity, exchangeability.
  - Zipf, Heap and other power laws

    (Pitman-Yao, 3-parameter IBP).
  - Flexible ways of building complex models

    (Hierarchical nonparametric models, dependent Dirichlet processes).

Are Nonparametric Models Nonparametric?

- Nonparametric just means *not parametric*: *cannot be described by a fixed set of parameters*.
  - Nonparametric models still have parameters, they just have an infinite number of them.

- No free lunch: *cannot learn from data unless you make assumptions*.
  - Nonparametric models still make modelling assumptions, they are just less constrained than the typical parametric models.

- Models can be nonparametric in one sense and parametric in another: **semiparametric** models.

Issues with Bayesian Nonparametrics

- Developing classes of nonparametric priors suitable for modelling data.
- Developing algorithms that can efficiently compute the posterior is important.
- Developing theory of asymptotics in nonparametric models.

Previous Tutorials and Reviews

- Mike Jordan's tutorial at NIPS 2005.
- Zoubin Ghahramani's tutorial at UAI 2005.
- Peter Orbanz' tutorial at MLSS 2009 (videolectures)
- My own tutorials at MLSS 2007, 2009 (videolectures), 2011 (Singapore) and elsewhere.
- Introduction to Dirichlet process [Teh 2010], nonparametric Bayes [Orbanz & Teh 2010, Gershman & Blei 2011], hierarchical Bayesian nonparametric models [Teh & Jordan 2010].
- Bayesian nonparametrics book [Hjort et al 2010].

- This tutorial: Dirichlet processes, Pitman-Yor processes, random partitions, random trees, hierarchical DPs, and hierarchical PYPs.

Dirichlet Process

Dirichlet Process

- Cornerstone of modern Bayesian nonparametrics.
- Rediscovered many times as the infinite limit of finite mixture models.
- Formally defined by [Ferguson 1973] as a distribution over measures.
- Can be derived in different ways, and as special cases of different processes.
- We will derive:
  - the infinite limit of a Gibbs sampler for finite mixture models
  - the Chinese restaurant process
  - the stick-breaking construction

The Infinite Limit of Finite Mixture Models

Finite Mixture Models

- Model for data from heterogeneous unknown sources.
- Each cluster (source) modelled using a parametric model (e.g. Gaussian).
- Data item *i*:

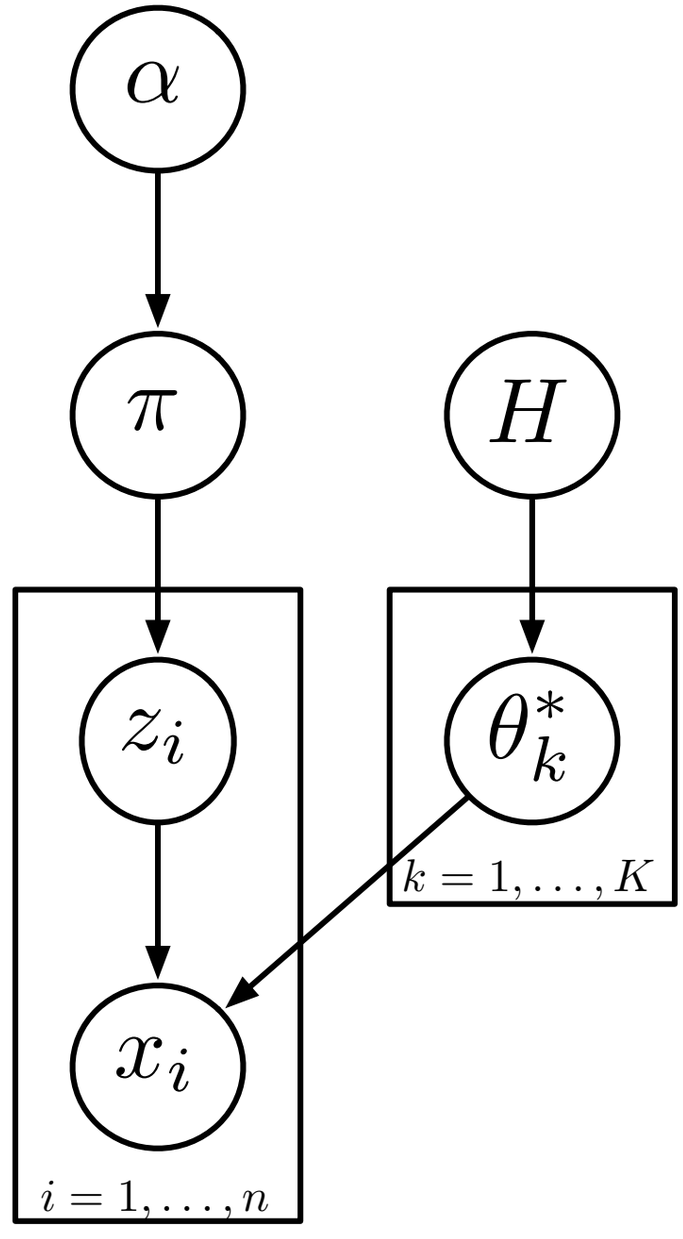
$$z_i|\pi \sim \text{Discrete}(\pi)$$
$$x_i|z_i, \theta^*_k \sim F(\theta^*_{z_i})$$

- **Mixing proportions**:

$$\pi = (\pi_1, \ldots, \pi_K)|\alpha \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

- Cluster *k*:

$$\theta^*_k|H \sim H$$



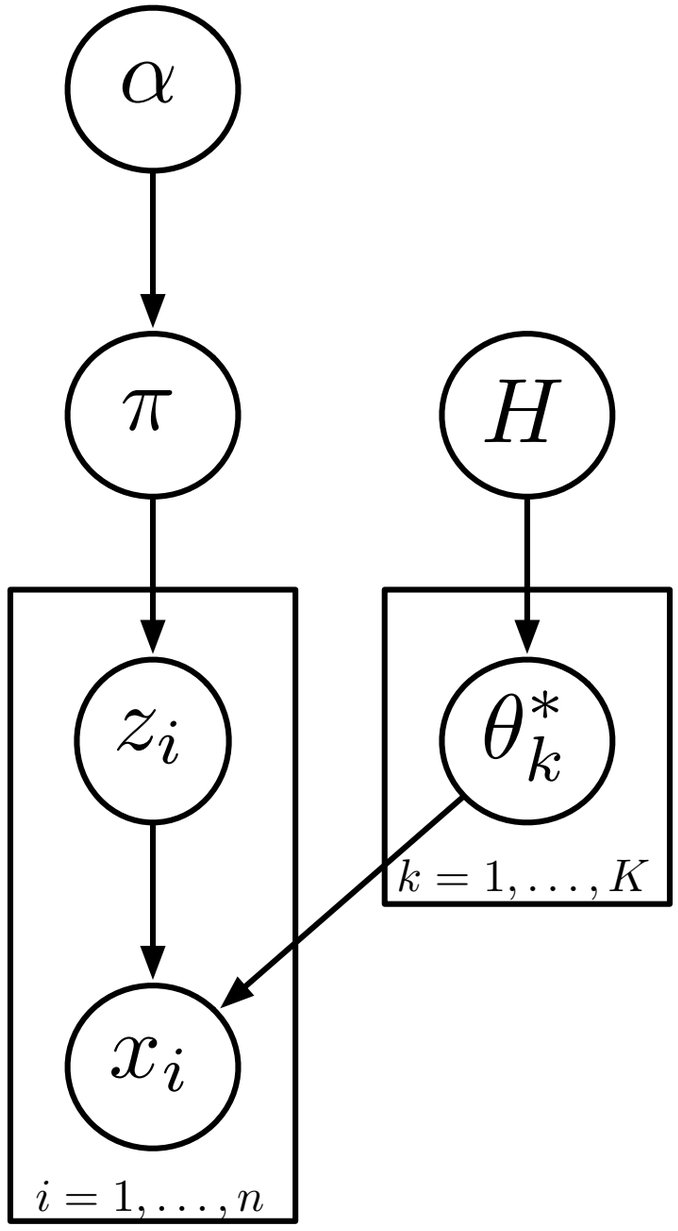
Finite Mixture Models

- Dirichlet distribution on the K-dimensional probability simplex $\{ \pi \mid \Sigma_k \pi_k = 1 \}$:

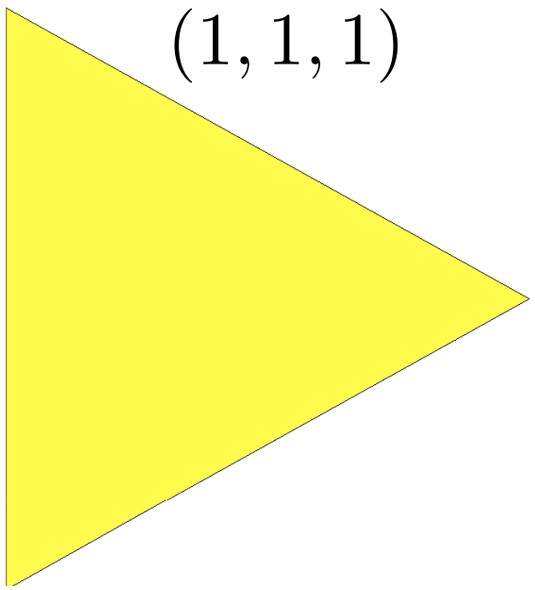
$$P(\pi|\alpha) = \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha/K)} \prod_{k=1}^{K} \pi_k^{\alpha/K-1}$$

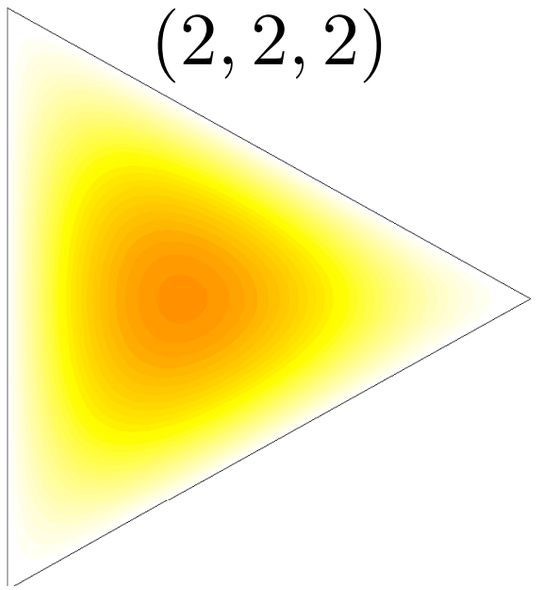
  with $\Gamma(a) = \int_0^\infty x^{a-1} e^x dx$.

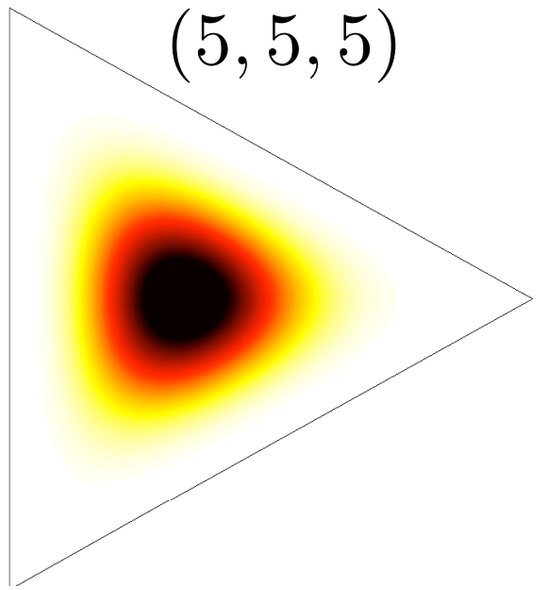
- Standard distribution on probability vectors, due to **conjugacy** with multinomial.

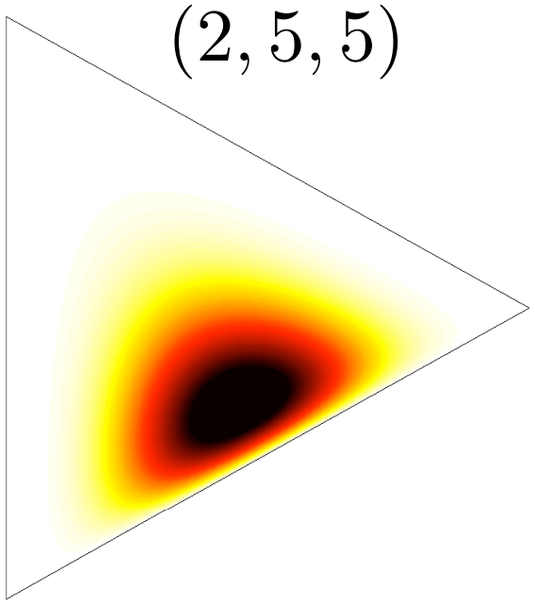


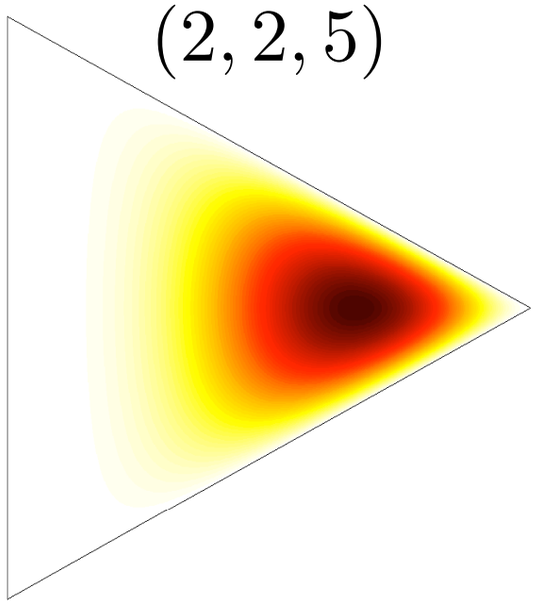
Dirichlet Distribution

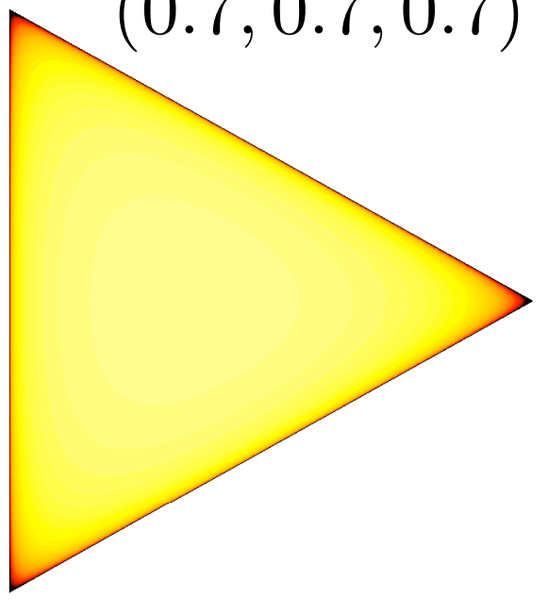












$$P(\pi|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{c=1}^{K} \pi_k^{\alpha_c - 1}$$

Dirichlet-Multinomial Conjugacy

- Joint distribution over z_i and π:

$$P(\pi|\alpha) \times \prod_{i=1}^{n} P(z_i|\pi) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K}\Gamma(\alpha/K)} \prod_{k=1}^{K} \pi_k^{\alpha/K-1} \times \prod_{k=1}^{K} \pi_k^{n_k}$$

  where $n_c = \#\{\, z_i = c \,\}$.

- Posterior distribution:

$$P(\pi|\mathbf{z},\alpha) = \frac{\Gamma(n+\alpha)}{\prod_{k=1}^{K}\Gamma(n_k+\alpha/K)} \prod_{k=1}^{K} \pi_k^{n_k+\alpha/K-1}$$

- Marginal distribution:

$$P(\mathbf{z}|\alpha) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K}\Gamma(\alpha/K)} \frac{\prod_{k=1}^{K}\Gamma(n_k+\alpha/K)}{\Gamma(n+\alpha)}$$

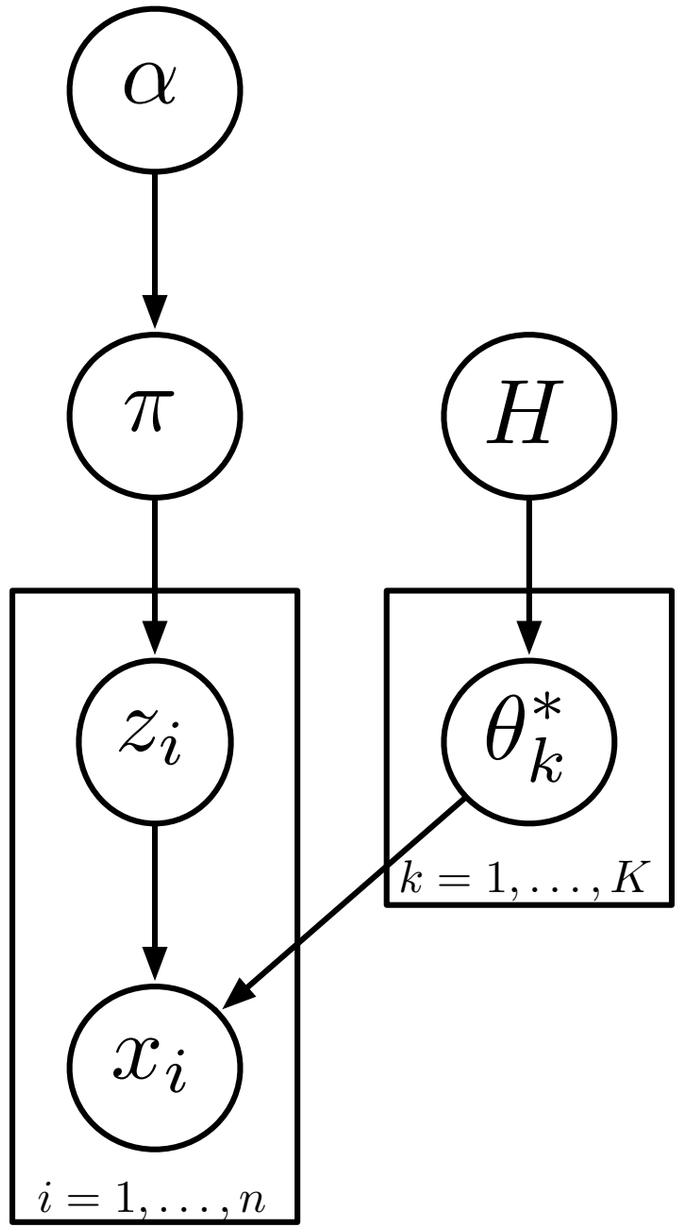
Gibbs Sampling

- All conditional distributions are simple to compute:

$$p(z_i = k|\text{others}) \propto \pi_k f(x_i|\theta_k^*)$$

$$\pi|\text{others} \sim \text{Dirichlet}(\tfrac{\alpha}{K} + n_1, \ldots, \tfrac{\alpha}{K} + n_K)$$

$$p(\theta_k^* = \theta|\text{others}) \propto h(\theta) \prod_{j:z_j=k} f(x_j|\theta)$$



- Not as efficient as collapsed Gibbs sampling, which integrates out π, θ*'s:

$$p(z_i = k|\text{others}) \propto \frac{\frac{\alpha}{K} + n_k^{\neg i}}{\alpha + n - 1} f(x_i|\{x_j : j \neq i, z_j = k\})$$

$$f(x_i|\{x_j : j \neq i, z_j = k\}) \propto \int h(\theta) f(x_i|\theta) \prod_{j \neq i: z_j = k} f(x_j|\theta) d\theta$$

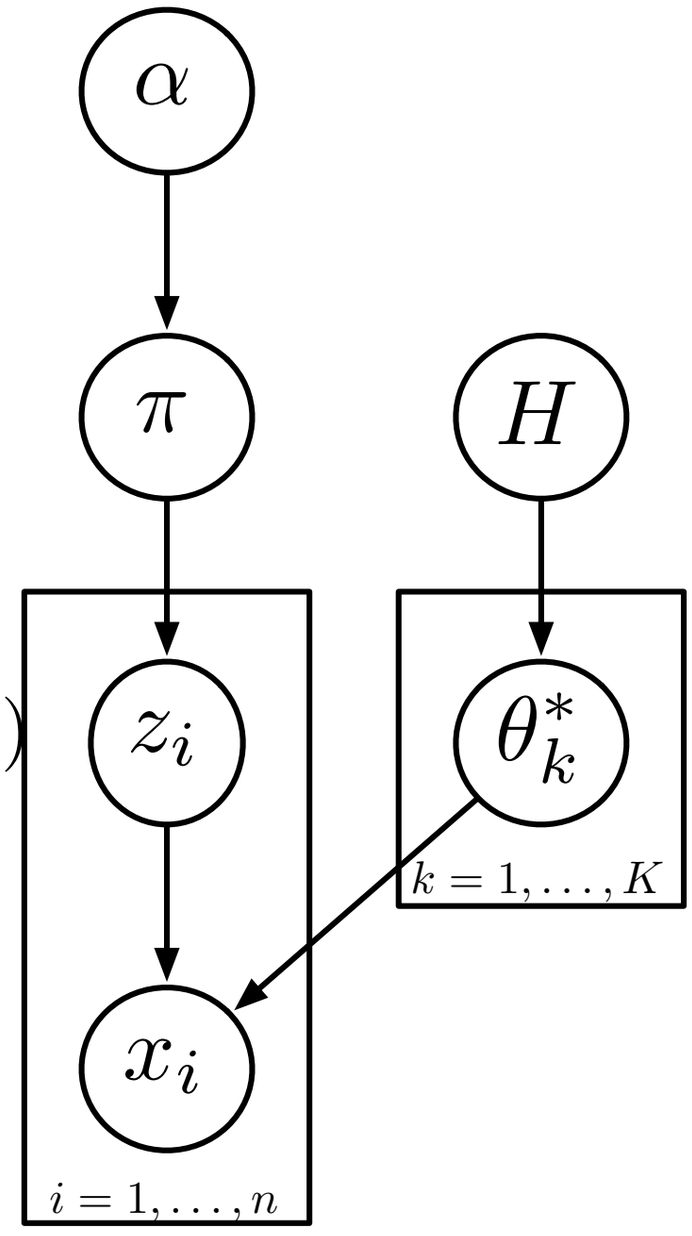
- Conditional distributions can be efficiently computed if *F* is conjugate to *H*.

Infinite Limit of Collapsed Gibbs Sampler

- We will take $K \rightarrow \infty$.
- Imagine a very large value of K.
- There are at most $n < K$ occupied clusters, so most components are empty. We can lump these empty components together:

$$p(z_i = k|\text{others}) = \frac{n_k^{\neg i} + \frac{\alpha}{K}}{n - 1 + \alpha} f(x_i|\{x_j : j \neq i, z_j = k\})$$

$$p(z_i = k_{\text{empty}}|\text{others}) = \frac{\alpha \frac{K - K^*}{K}}{n - 1 + \alpha} f(x_i|\{\})$$



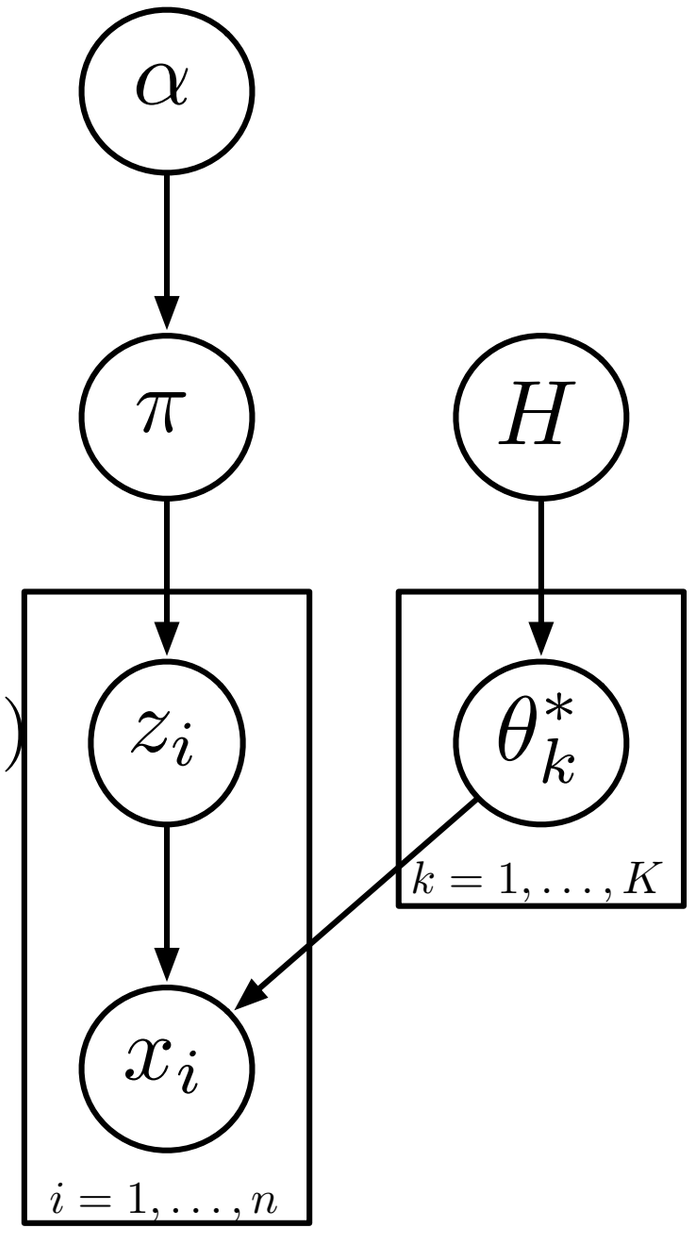
[Neal 2000, Rasmussen 2000, Ishwaran & Zarepour 2002]

Infinite Limit of Collapsed Gibbs Sampler

- We will take $K \rightarrow \infty$.
- Imagine a very large value of K.
- There are at most $n < K$ occupied clusters, so most components are empty. We can lump these empty components together:

$$p(z_i = k|\text{others}) = \frac{n_k^{\neg i}}{n-1+\alpha} f(x_i|\{x_j : j \neq i, z_j = k\})$$

$$p(z_i = k_{\text{empty}}|\text{others}) = \frac{\alpha}{n-1+\alpha} f(x_i|\{\})$$



[Neal 2000, Rasmussen 2000, Ishwaran & Zarepour 2002]

Infinite Limit

- The actual infinite limit of the finite mixture model does not make sense:
  - any particular cluster will get a mixing proportion of 0.
- Better ways of making this infinite limit precise:
  - Chinese restaurant process.
  - Stick-breaking construction.
- Both are different views of the Dirichlet process (DP).
- DPs can be thought of as infinite dimensional Dirichlet distributions.
- The $K \rightarrow \infty$ Gibbs sampler is for DP mixture models.

Ferguson's Definition of the Dirichlet Process

Tiny Bit of Probability Theory

- A σ-**algebra** Σ is a family of subsets of a set Θ such that
  - Σ is not empty;
  - if $A \in \Sigma$ then $\Theta \backslash A \in \Sigma$;
  - if $A_1, A_2, \ldots \in \Sigma$ then $\cup_i A_i \in \Sigma$.
- (Θ, Σ) is a **measure space** and $A \in \Sigma$ are the **measurable sets**.
- A **measure** μ over (Θ, Σ) is a function $\mu : \Sigma \rightarrow [0, \infty]$ such that
  - $\mu(\emptyset) = 0$;
  - if $A_1, A_2, \ldots \in \Sigma$ are disjoint then $\mu(\cup_i A_i) = \Sigma_i \mu(A_i)$;
  - a **probability measure** is one where $\mu(\Theta) = 1$.
- Everything we consider here will be measurable.

Tiny Bit of Probability Theory

- Given two measure spaces (Θ, Σ) and (Δ, Φ) a function $f : \Theta \to \Delta$ is **measurable** if $f^{-1}(A) \in \Sigma$ for every $A \in \Phi$.
- If P is a probability measure on (Θ, Σ), a **random variable** X taking values in Δ is simply a measurable function $X : \Theta \to \Delta$.
  - This of the probability space (Θ, Σ, P) as a black-box random number generator, and X as a fixed function taking random samples in Θ and producing random samples in Δ.
  - The probability of an event $A \in \Phi$ is $P(X \in A) = P(X^{-1}(A))$.
- A **stochastic process** is simply a collection of random variables $\{X_i\}_{i \in I}$ over the same measure space (Θ, Σ), where I is an index set.
  - I can be an infinite (even uncountably infinite) set.

Ferguson's Definition of Dirichlet Processes

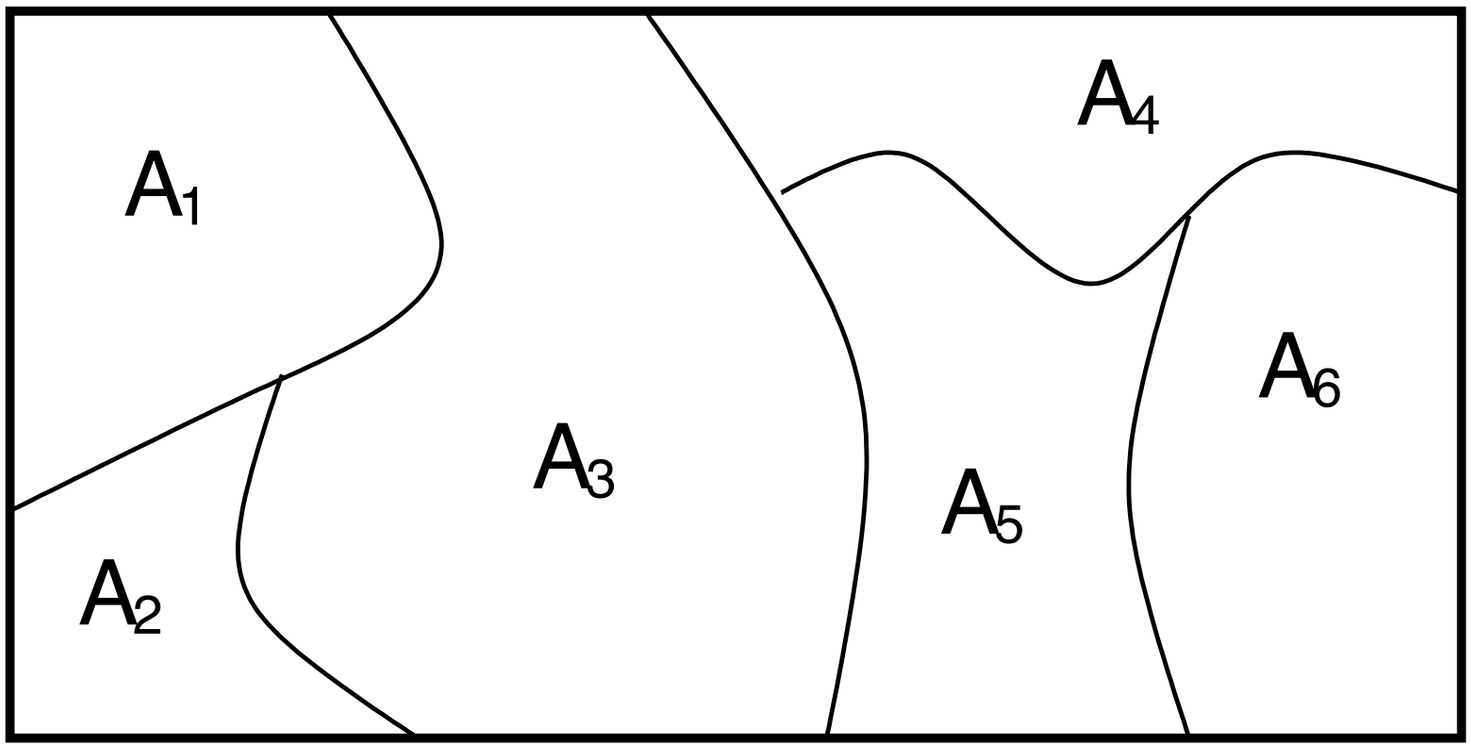
- A **Dirichlet process** (DP) is a random probability measure G over (Θ, Σ) such that for any finite set of measurable sets $A_1,\ldots A_K \in \Sigma$ partitioning Θ, i.e.

$$A_1 \dot{\cup} \cdots \dot{\cup} A_K = \Theta$$

  we have

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$

where α and H are parameters of the DP.



[Ferguson 1973]

Parameters of the Dirichlet Process

- α is called the **strength, mass** or **concentration parameter**.
- H is called the **base distribution**.
- Mean and variance:

$$\mathbb{E}[G(A)] = H(A)$$

$$\mathbb{V}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

  where A is a measurable subset of Θ.

- H is the mean of G, and α is an inverse variance.

Posterior Dirichlet Process

- Suppose

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

- We can define random variables that are G distributed:

$$\theta_i | G \sim G \quad \text{for } i = 1, \ldots, n$$

- The usual Dirichlet-multinomial conjugacy carries over to the DP as well:

$$G | \theta_1, \ldots, \theta_n \sim \mathrm{DP}(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n})$$

Pólya Urn Scheme

$$G \sim \mathrm{DP}(\alpha, H)$$
$$\theta_i | G \sim G \quad \text{for } i = 1, 2, \ldots$$

- Marginalizing out G, we get:

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$$

- This is called the **Pólya**, **Hoppe** or **Blackwell-MacQueen urn scheme**.
  - Start with an urn with α balls of a special colour.
  - Pick a ball randomly from urn:
    - If it is a special colour, make a new ball with colour sampled from H, note the colour, and return both balls to urn.
    - If not, note its colour and return two balls of that colour to urn.

[Blackwell & MacQueen 1973, Hoppe 1984]

Clustering Property

$$G \sim \mathrm{DP}(\alpha, H)$$
$$\theta_i | G \sim G \quad \text{for } i = 1, 2, \ldots$$

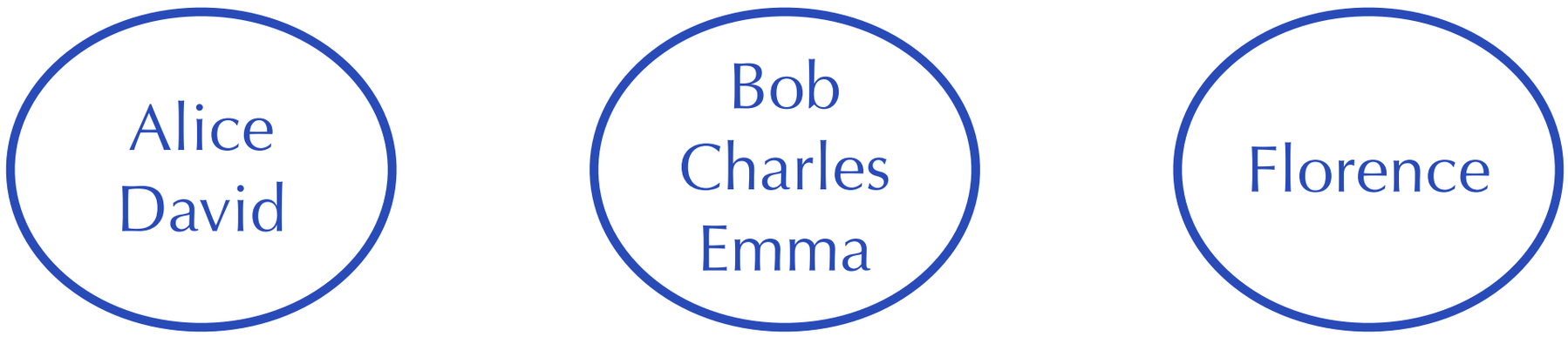
- The n variables $\theta_1, \theta_2, \ldots, \theta_n$ can take on $K \leq n$ distinct values.
- Let the distinct values be $\theta_1^*, \ldots, \theta_K^*$. This defines a partition of $\{1, \ldots, n\}$ such that i is in cluster k if and only if $\theta_i = \theta_k^*$.
- The induced distribution over partitions is the **Chinese restaurant process**.

Chinese Restaurant Process

[Aldous 1985, Pitman 2006]

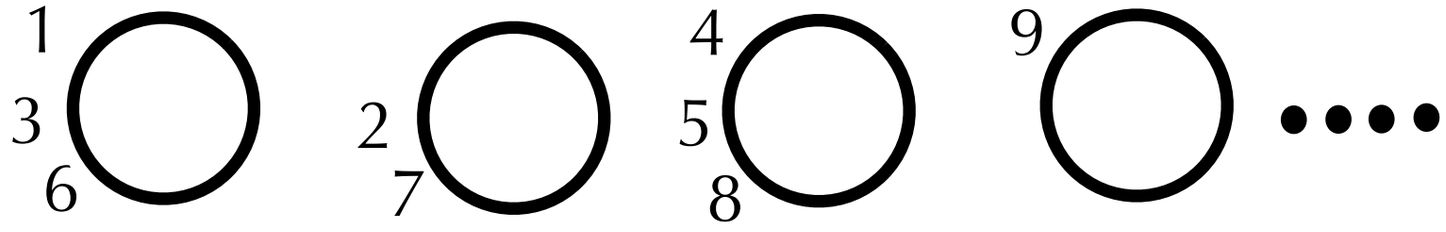
Partitions

- A **partition** ϱ of a set S is:
  - A disjoint family of non-empty subsets of S whose union in S.
  - S = {Alice, Bob, Charles, David, Emma, Florence}.
  - ϱ = { {Alice, David}, {Bob, Charles, Emma}, {Florence} }.



- Denote the set of all partitions of S as $\mathcal{P}_S$.
- **Random partitions** are random variables taking values in $\mathcal{P}_S$.
- We will work with partitions of $S = [n] = \{1,2,...n\}$.

Chinese Restaurant Process



- Each customer comes into restaurant and sits at a table:

$$p(\text{sit at table } c) = \frac{n_c}{\alpha + \sum_{c \in \varrho} n_c} \qquad p(\text{sit at new table}) = \frac{\alpha}{\alpha + \sum_{c \in \varrho} n_c}$$

- Customers correspond to elements of S, and tables to clusters in ϱ.
- **Rich-gets-richer**: large clusters more likely to attract more customers.
- Multiplying conditional probabilities together, the overall probability of ϱ, called the **exchangeable partition probability function** (EPPF), is:

$$P(\varrho|\alpha) = \frac{\alpha^{|\varrho|}\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{c \in \varrho} \Gamma(|c|)$$

[Aldous 1985, Pitman 2006]

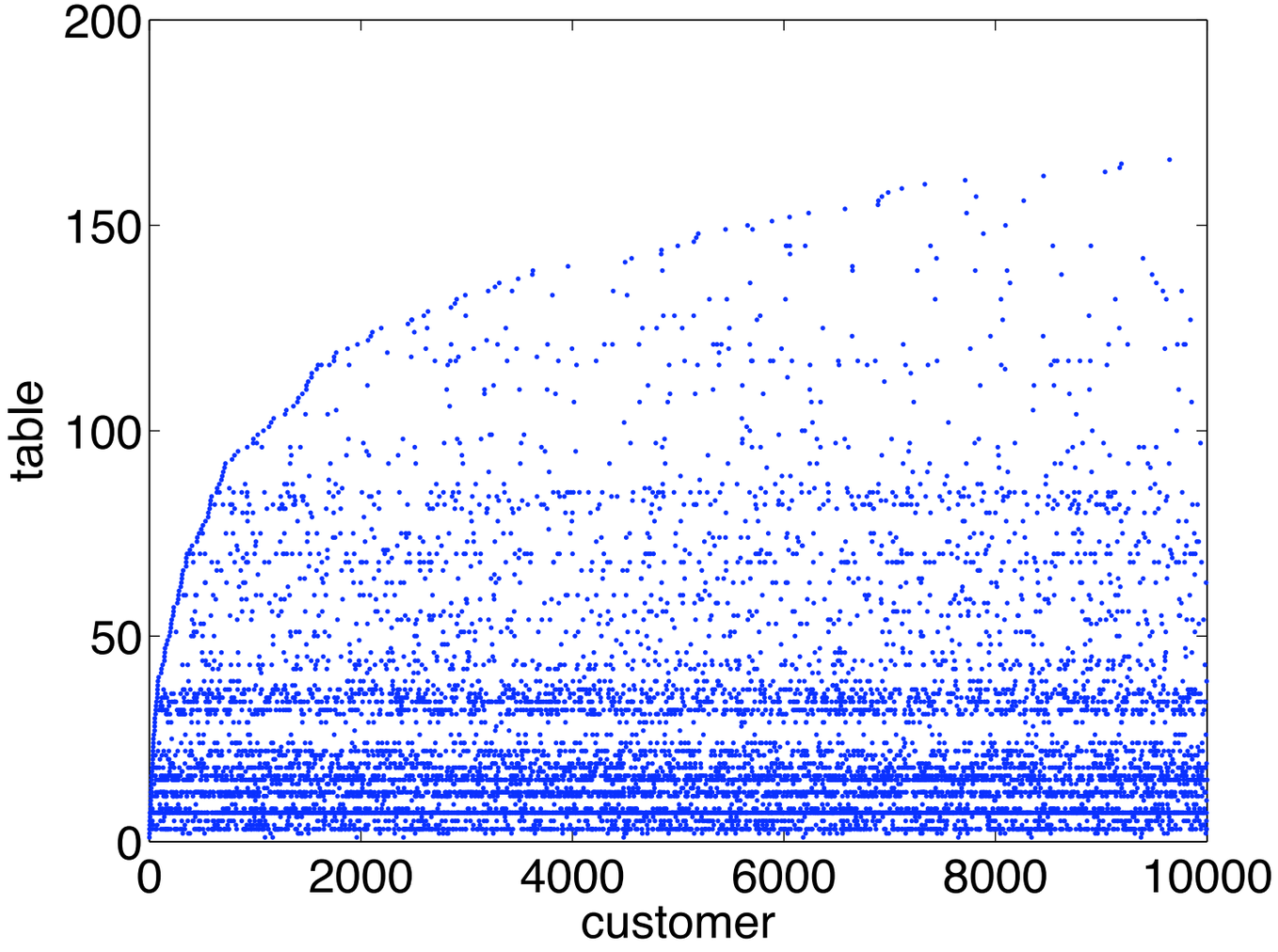
Number of Clusters

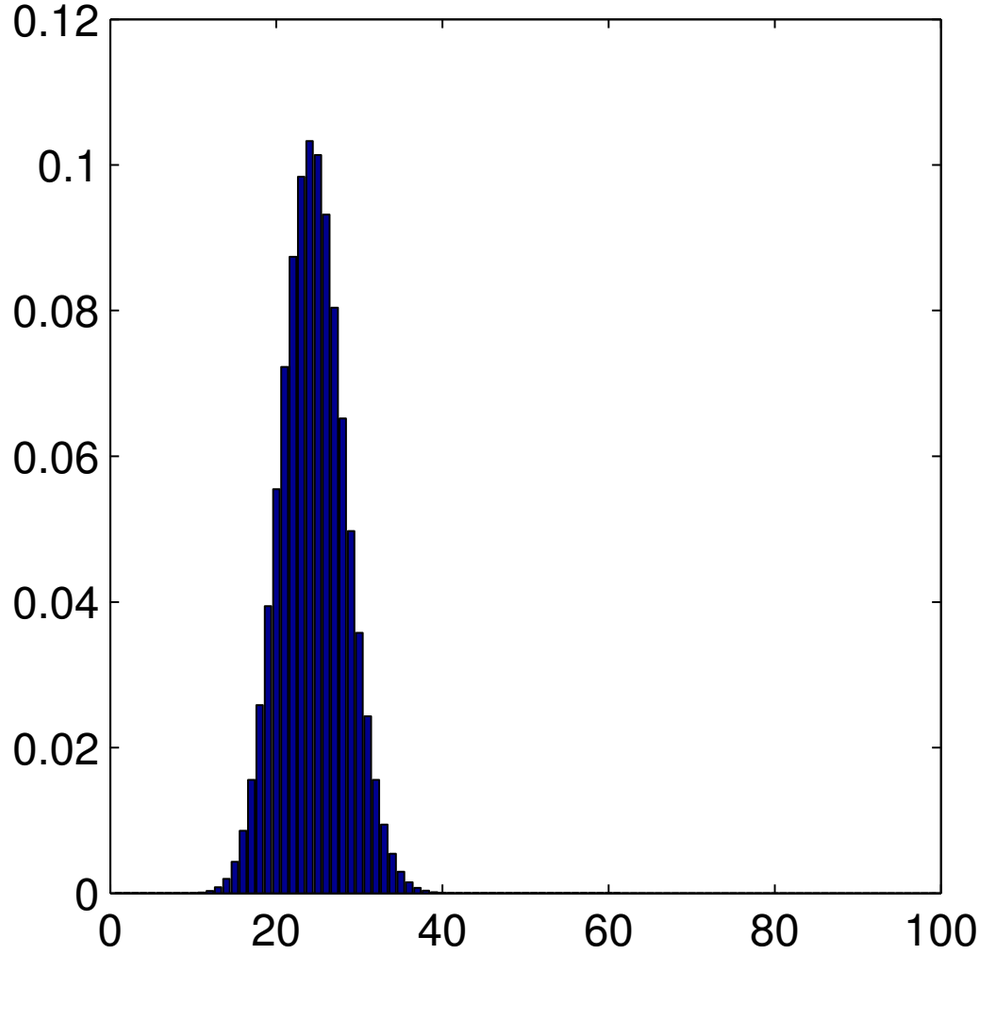
- The prior mean and variance of K are:

$$\mathbb{E}[||\rho||\alpha, n] = \alpha(\psi(\alpha + n) - \psi(\alpha)) \approx \alpha \log\left(1 + \frac{n}{\alpha}\right)$$

$$\mathbb{V}[||\rho||\alpha, n] = \alpha(\psi(\alpha + n) - \psi(\alpha)) + \alpha^2(\psi'(\alpha + n) - \psi'(\alpha)) \approx \alpha \log\left(1 + \frac{n}{\alpha}\right)$$

$$\psi(\alpha) = \frac{\partial}{\partial \alpha} \log \Gamma(\alpha)$$





Model-based Clustering with Chinese Restaurant Process

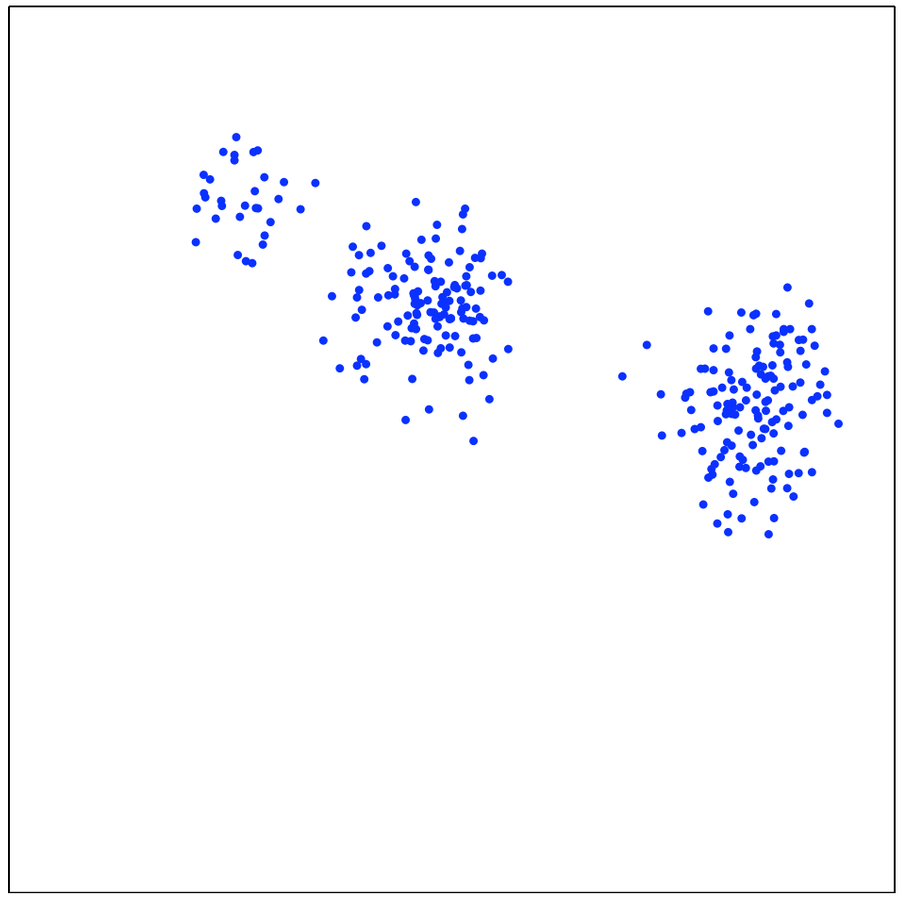
Partitions in Model-based Clustering

- Partitions are the natural latent objects of inference in clustering.
    - Given a dataset S, partition it into clusters of similar items.
- Cluster $c \in \varrho$ described by a model

$$F(\theta_c^*)$$

parameterized by θ_c^*.

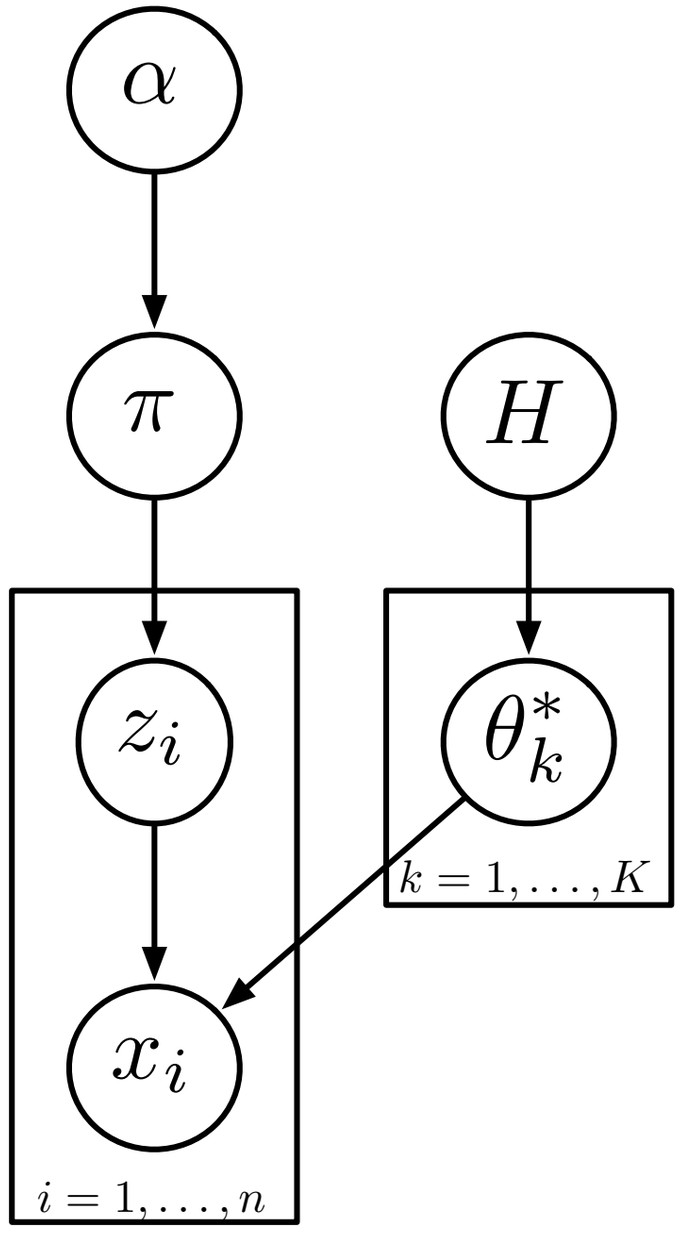
- Bayesian approach: introduce prior over ϱ and θ_c^*; compute posterior over both.



Finite Mixture Model

- Explicitly allow only K clusters in partition:
  - Each cluster k has parameter θ_k.
  - Each data item i assigned to k with **mixing probability** π_k.
    - Gives a random partition with at most K clusters.
- Priors on the other parameters:

$$\pi|\alpha \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$
$$\theta_k^*|H \sim H$$



Induced Distribution over Partitions

$$P(\mathbf{z}|\alpha) = \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha/K)} \frac{\prod_k \Gamma(n_k + \alpha/K)}{\Gamma(n + \alpha)}$$

- P($\mathbf{z}|\alpha$) describes a partition of the data set into clusters, *and a labelling of each cluster with a mixture component index*.
- Induces a distribution over partitions ϱ (without labelling) of the data set:

$$P(\varrho|\alpha) = [K]^k_{-1} \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{c \in \varrho} \frac{\Gamma(|c| + \alpha/K)}{\Gamma(\alpha/K)}$$

  where $[x]^a_b = x(x + b) \cdots (x + (a - 1)b)$.

- Taking $K \to \infty$, we get a proper distribution over partitions without a limit on the number of clusters:

$$P(\varrho|\alpha) \to \frac{\alpha^{|\varrho|}\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{c \in \varrho} \Gamma(|c|)$$

Chinese Restaurant Process

- An important representation of the Dirichlet process
- An important object of study in its own right.
- Predates the Dirichlet process and originated in genetics (related to Ewen's sampling formula there).
- Large number of MCMC samplers using CRP representation.
- Random partitions are useful concepts for clustering problems in machine learning
  - CRP mixture models for nonparametric model-based clustering.
  - hierarchical clustering using concepts of fragmentations and coagulations.
  - clustering nodes in graphs, e.g. for community discovery in social nets.
  - Other combinatorial structures can be built from partitions.

Stick-breaking Construction

Clustering Property

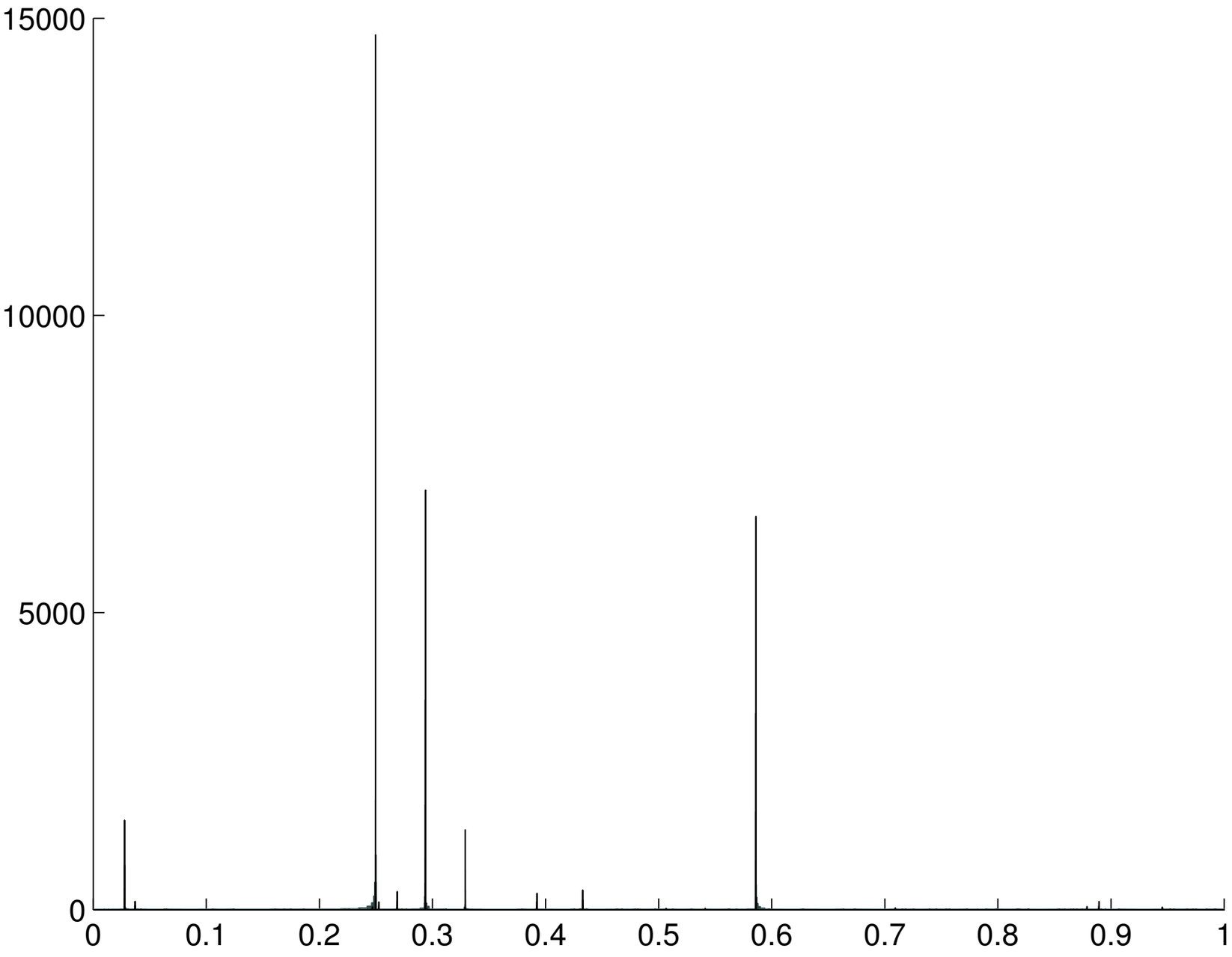
$$G \sim \mathrm{DP}(\alpha, H)$$
$$\theta_i | G \sim G \quad \text{for } i = 1, 2, \ldots$$

- The same values can be repeated among the variables $\theta_1, \theta_2, \ldots, \theta_n$.

- This can only be the case if G is an atomic distribution.

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

A draw from a Dirichlet Process



Atomic Distributions

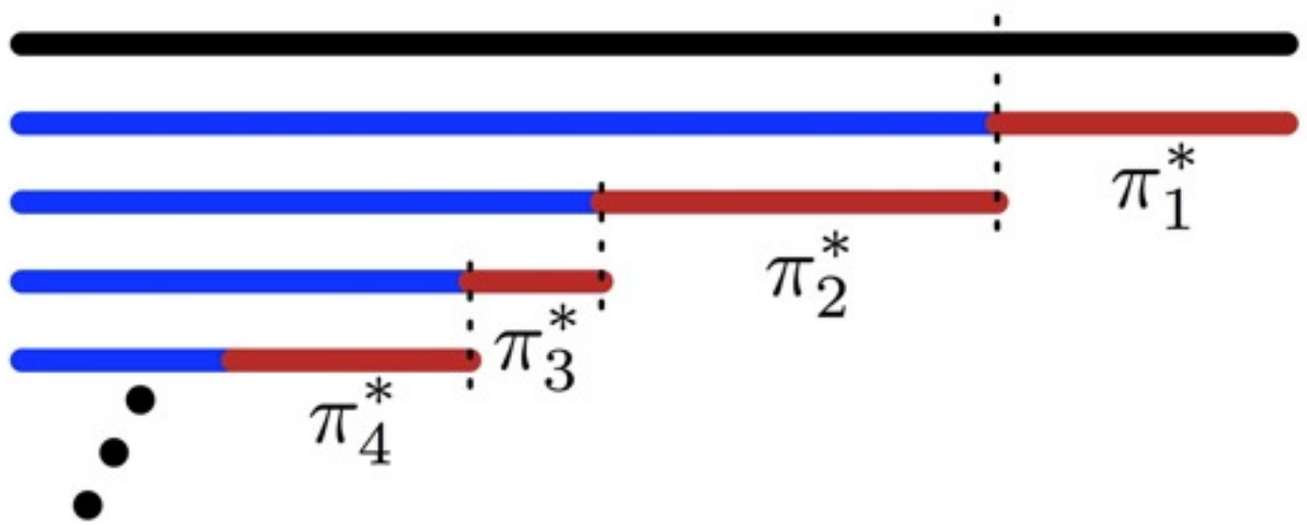
- Draws from Dirichlet processes will always be atomic:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

where $\Sigma_k \pi_k = 1$ and $\theta_k^* \in \Theta$.

- A number of ways to specify the joint distribution of $\{\pi_k, \theta_k^*\}$.
  - Stick-breaking construction;
  - Poisson-Dirichlet distribution.

Stick-breaking Construction



- **Stick-breaking construction** for the joint distribution:

$$\theta_k^* \sim H \qquad v_k \sim \text{Beta}(1, \alpha) \qquad \text{for } k = 1, 2, \ldots.$$

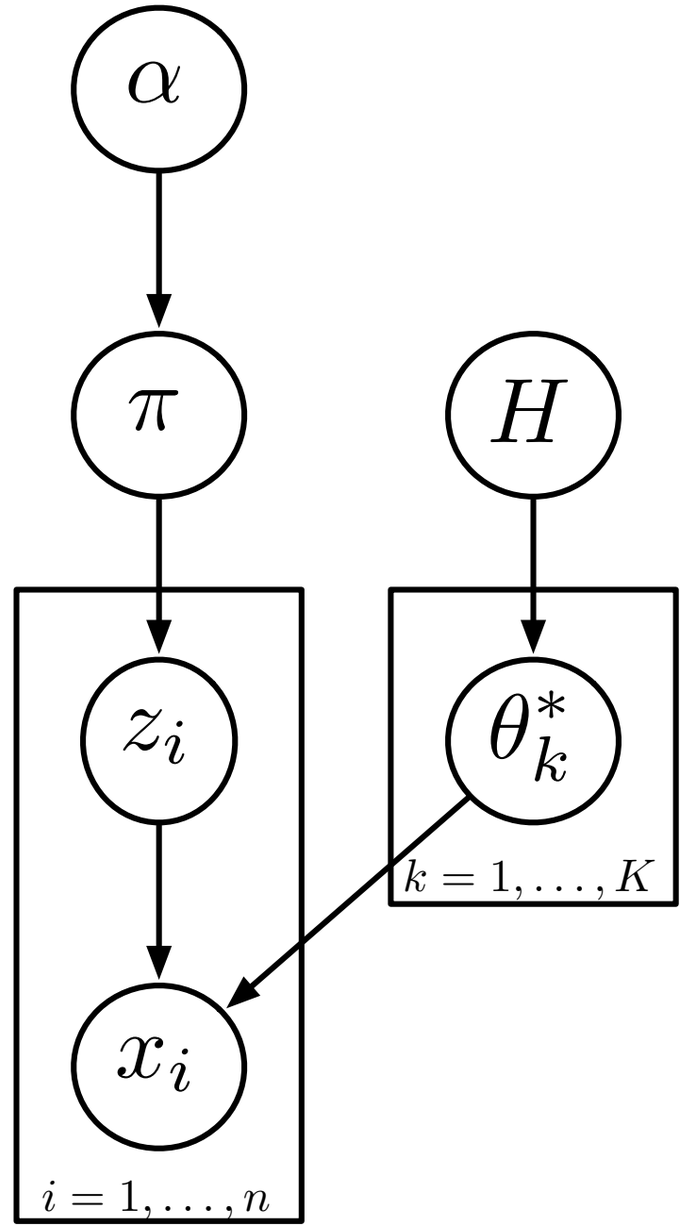
$$\pi_k^* = v_k \prod_{j=1}^{k-1} (1 - v_j) \qquad G = \sum_{k=1}^{\infty} \pi_k^* \delta_{\theta_k^*}$$

- π_k's are decreasing on average but not strictly.
- Distribution of $\{\pi_k\}$ is the **Griffiths-Engen-McCloskey** (GEM) distribution.
- **Poisson-Dirichlet distribution** [Kingman 1975] gives a strictly decreasing ordering (but is not computationally tractable).

Finite Mixture Model

- Explicitly allow only K clusters in partition:
  - Each cluster k has parameter θ_k.
  - Each data item i assigned to k with mixing probability π_k.
  - Gives a random partition with at most K clusters.
- Priors on the other parameters:

$$\pi|\alpha \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$
$$\theta_k^*|H \sim H$$



Size-biased Permutation

- Reordering clusters do not change the marginal distribution on partitions or data items.
- By strictly decreasing π_k: Poisson-Dirichlet distribution.
- Reorder stochastically as follows gives stick-breaking construction:
  - Pick cluster k to be first cluster with probability π_k .
  - Remove cluster k and renormalize rest of $\{ \pi_k : j \neq k \}$; repeat.
- Stochastic reordering is called a **size-biased permutation**.
- After reordering, taking $K \to \infty$ gives the corresponding DP representations.

Stick-breaking Construction

- Easy to generalize stick-breaking construction:
  - to other random measures;
  - to random measures that depend on covariates or vary spatially.
- Easy to work with different algorithms:
  - MCMC samplers;
  - variational inference;
  - parallelized algorithms.

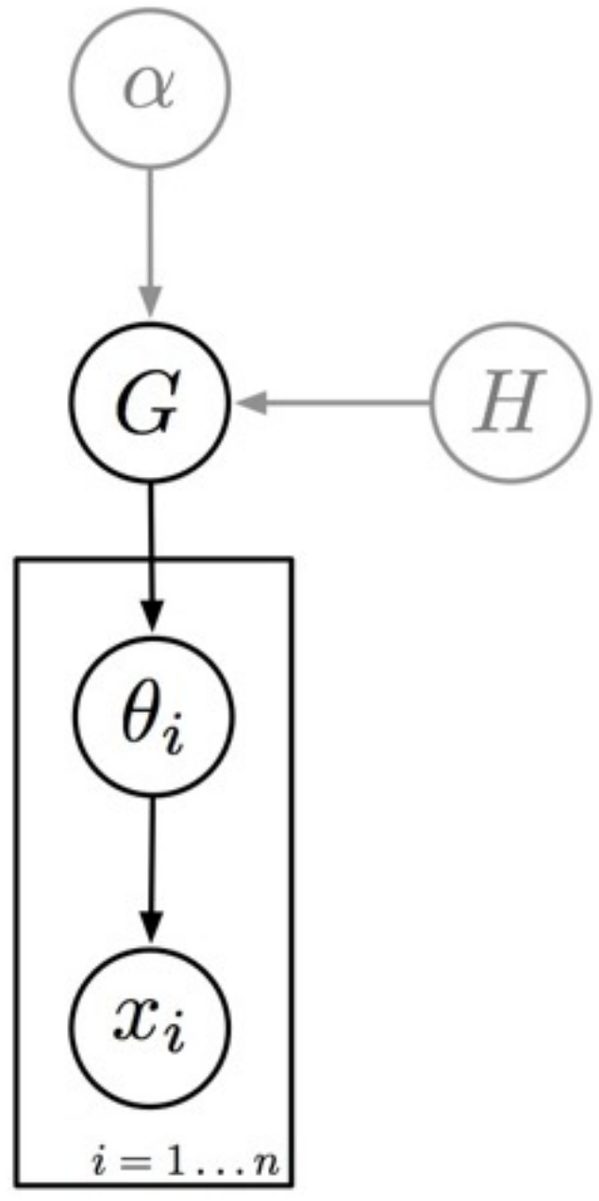
[Ishwaran & James 2001, Dunson 2010 and many others]

DP Mixture Model: Representations and Inference

DP Mixture Model

- A **DP mixture model**:

$$G|\alpha, H \sim \mathrm{DP}(\alpha, H)$$
$$\theta_i|G \sim G$$
$$x_i|\theta_i \sim F(\theta_i)$$



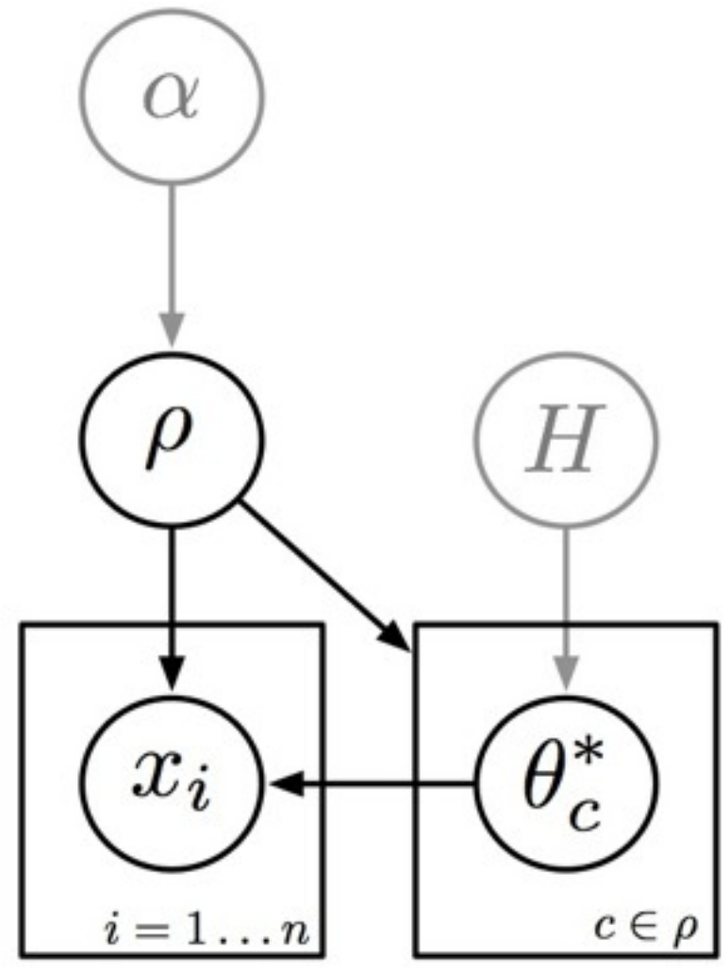
- Different representations:
  - $\theta_1, \theta_2, \ldots, \theta_n$ are clustered according to Pólya urn scheme, with induced partition given by a CRP.
  - G is atomic with weights and atoms described by stick-breaking construction.

[Neal 2000, Rasmussen 2000, Ishwaran & Zarepour 2002]

CRP Representation

- Representing the partition structure explicitly with a CRP:

$$\rho|\alpha \sim \mathrm{CRP}([n], \alpha)$$
$$\theta_c^*|H \sim H \text{ for } c \in \rho$$
$$x_i|\theta_c^* \sim F(\theta_c^*) \text{ for } c \ni i$$



- Makes explicit that this is a clustering model.
- Using a CRP prior for ϱ obviates need to limit number of clusters as in finite mixture models.

[Neal 2000, Rasmussen 2000, Ishwaran & Zarepour 2002]

Marginal Sampler

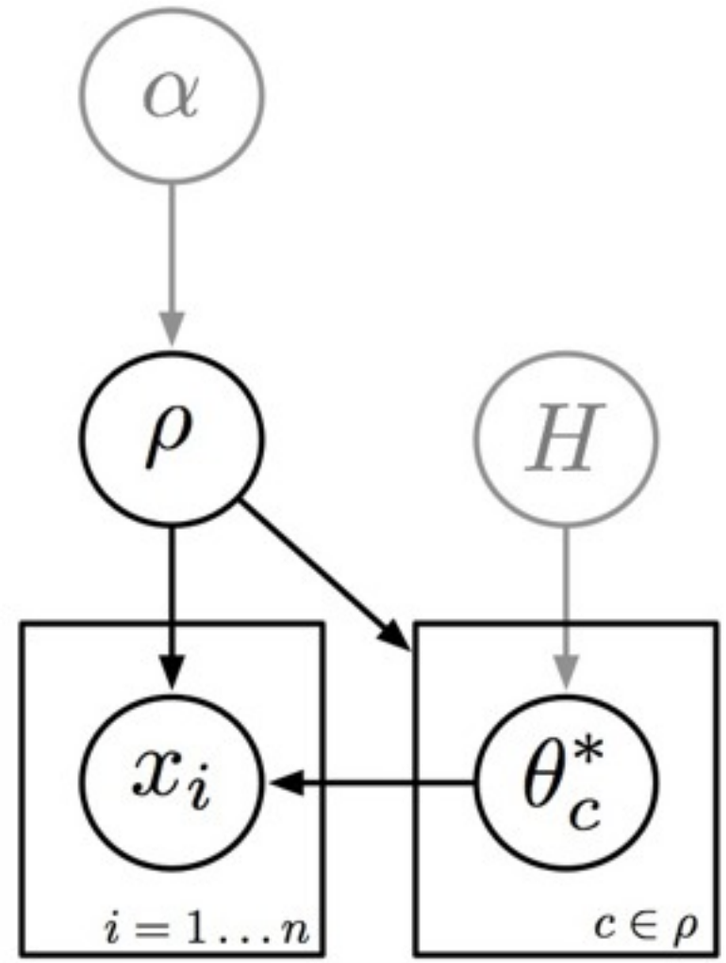
$$\rho|\alpha \sim \mathrm{CRP}([n], \alpha)$$
$$\theta_c^*|H \sim H \text{ for } c \in \rho$$
$$x_i|\theta_c^* \sim F(\theta_c^*) \text{ for } c \ni i$$

- "Marginal" MCMC sampler.
  - Marginalize out *G*, and Gibbs sample partition.
- Conditional probability of cluster of data item *i*:

$$P(\rho_i|\rho_{\setminus i}, \mathbf{x}, \boldsymbol{\theta}) = P(\rho_i|\rho_{\setminus i}) P(x_i|\rho_i, \mathbf{x}_{\setminus i}, \boldsymbol{\theta})$$

$$P(\rho_i|\rho_{\setminus i}) = \begin{cases} \frac{|c|}{n-1+\alpha} & \text{if } \rho_i = c \in \rho_{\setminus i} \\ \frac{\alpha}{n-1+\alpha} & \text{if } \rho_i = \text{new} \end{cases}$$

$$P(x_i|\rho_i, \mathbf{x}_{\setminus i}, \boldsymbol{\theta}) = \begin{cases} f(x_i|\theta_{\rho_i}) & \text{if } \rho_i = c \in \rho_{\setminus i} \\ \int f(x_i|\theta) h(\theta) d\theta & \text{if } \rho_i = \text{new} \end{cases}$$



- A variety of methods to deal with new clusters.
- Difficulty lies in dealing with new clusters, especially when prior *h* is not conjugate to *f*.

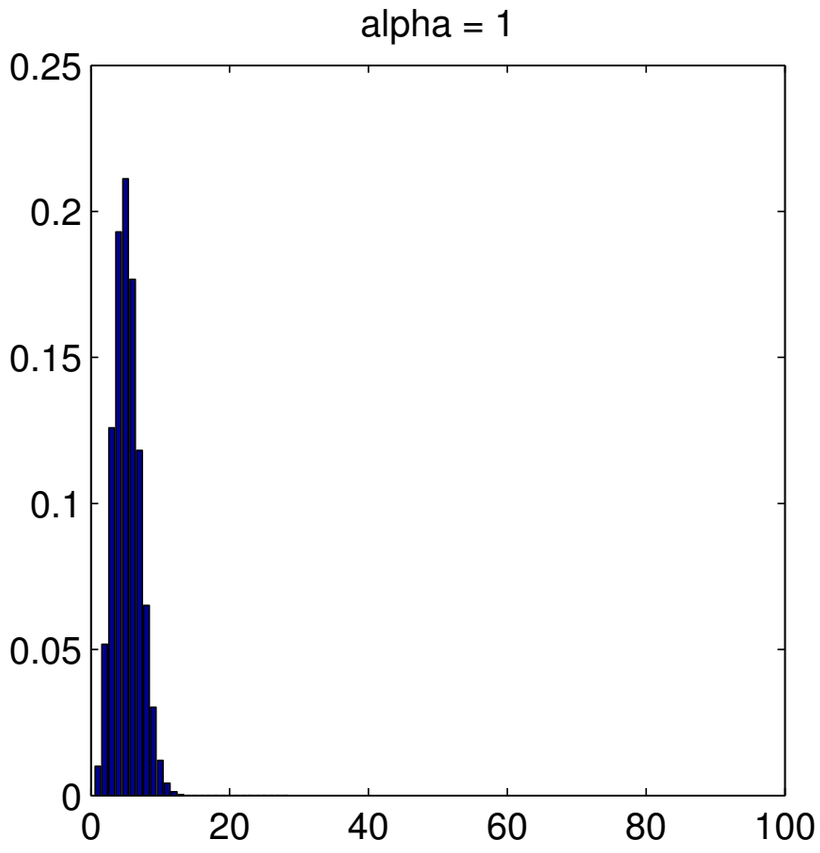
[Neal 2000]

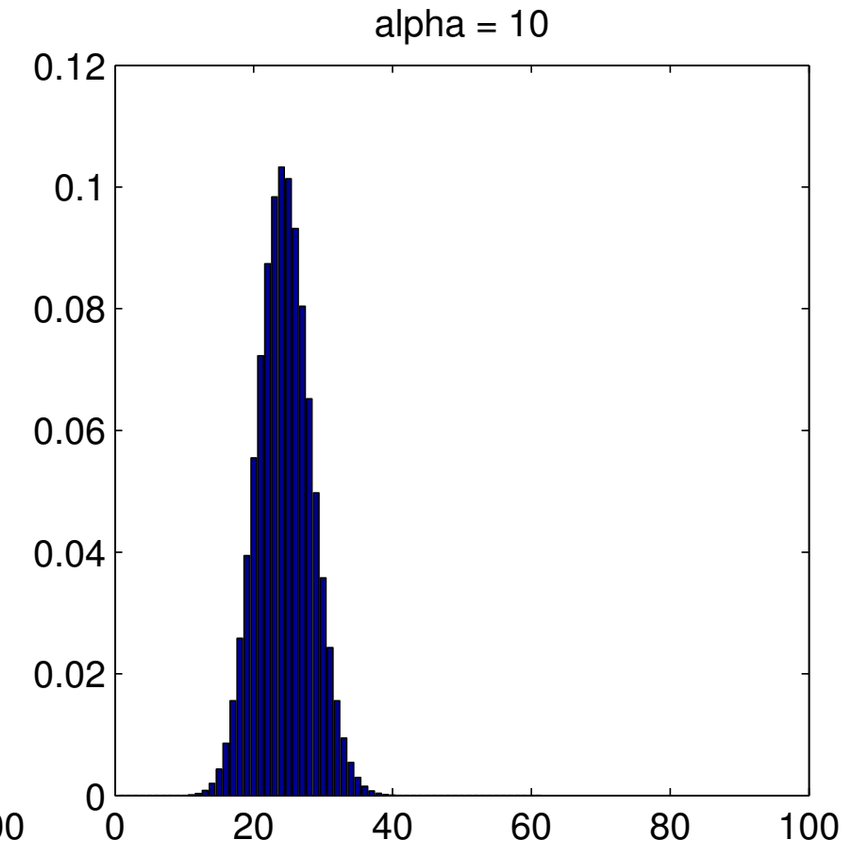
Induced Prior on the Number of Clusters

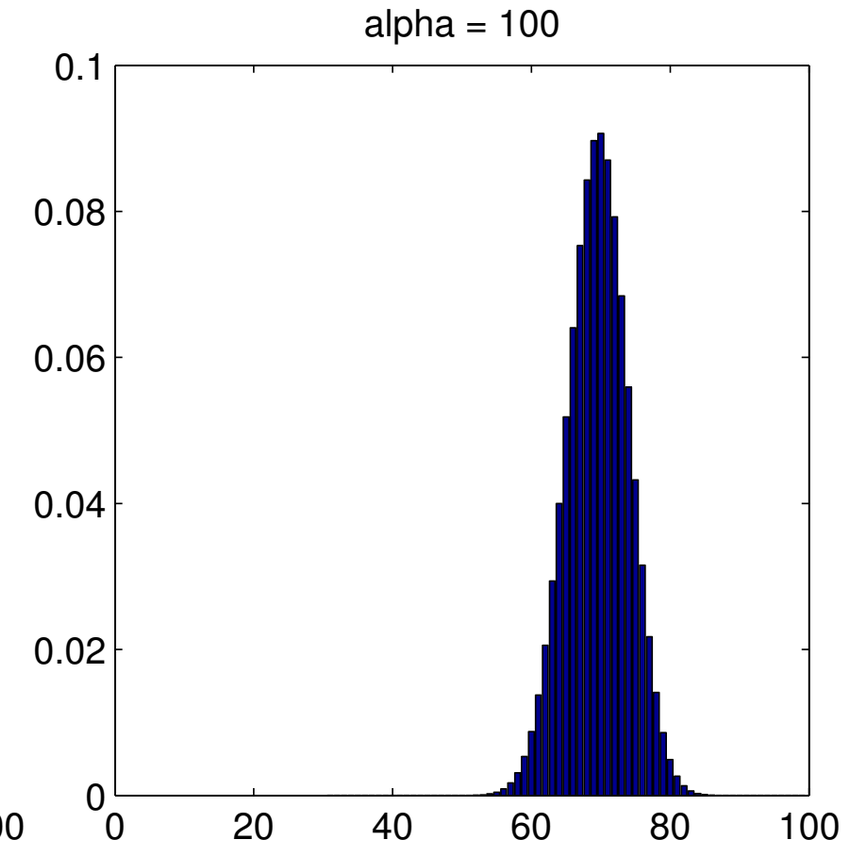
- The prior expectation and variance of $|\varrho|$ are:

$$\mathbb{E}[|\rho||\alpha, n] = \alpha(\psi(\alpha + n) - \psi(\alpha)) \approx \alpha \log\left(1 + \frac{n}{\alpha}\right)$$

$$\mathbb{V}[|\rho||\alpha, n] = \alpha(\psi(\alpha + n) - \psi(\alpha)) + \alpha^2(\psi'(\alpha + n) - \psi'(\alpha)) \approx \alpha \log\left(1 + \frac{n}{\alpha}\right)$$



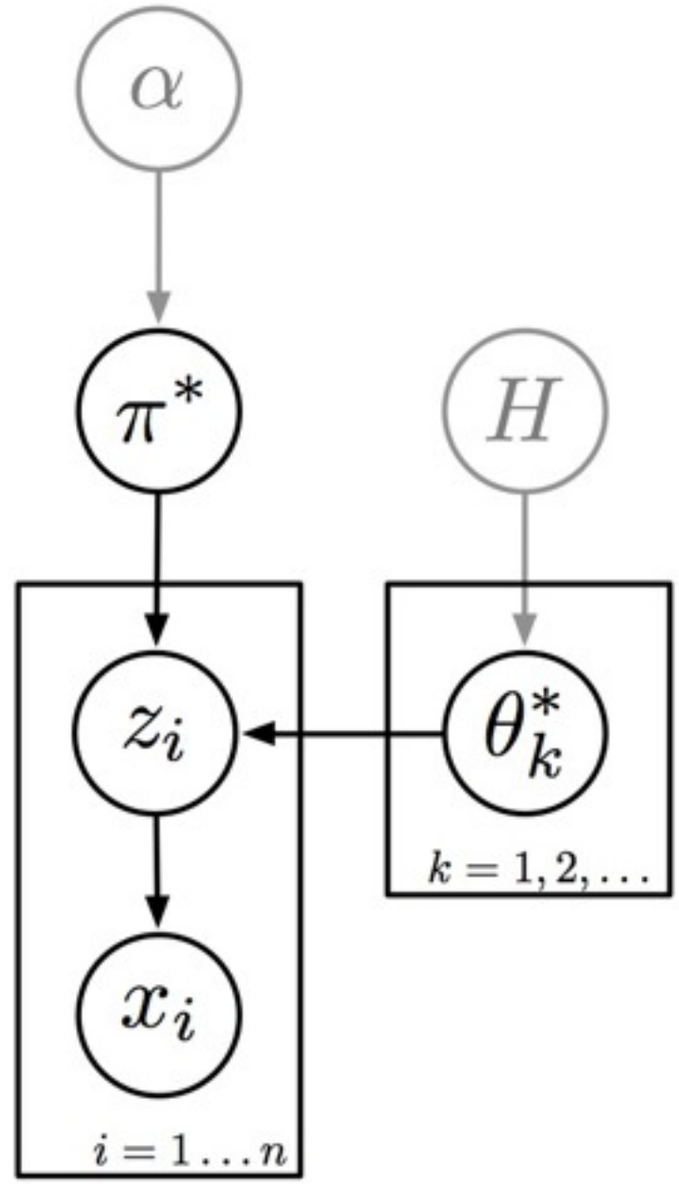




Stick-breaking Representation

- Dissecting stick-breaking representation for G:

$$\pi^*|\alpha \sim \text{GEM}(\alpha)$$
$$\theta_k^*|H \sim H$$
$$z_i|\pi^* \sim \text{Discrete}(\pi^*)$$
$$x_i|z_i, \theta_{z_i}^* \sim F(\theta_{z_i}^*)$$



- Makes explicit that this is a mixture model with an infinite number of components.
- Conditional sampler:
  - Standard Gibbs sampler, except need to truncate the number of clusters.
  - Easy to work with non-conjugate priors.
  - For sampler to mix well need to introduce moves for permuting the order of clusters.

[Ishwaran & James 2001, Walker 2007, Papaspiliopoulos & Roberts 2008]

Explicit G Sampler

- Represent G explicitly, alternately sampling $\{\theta_i\}|G$ (simple) and $G|\{\theta_i\}$:.

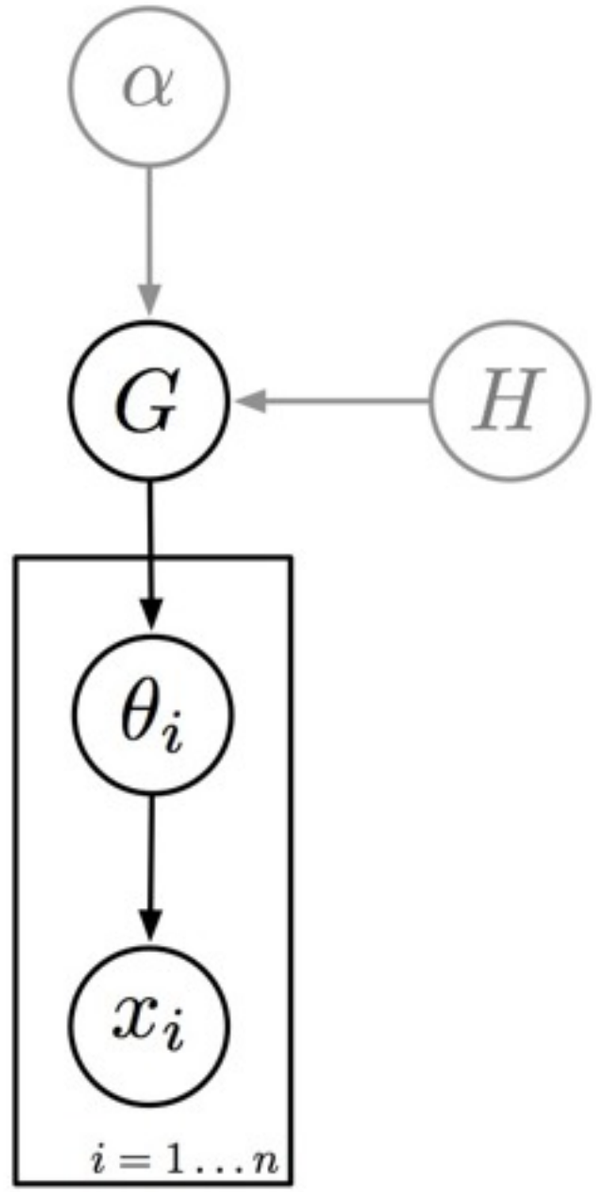
$$G|\theta_1, \ldots, \theta_n \sim \mathrm{DP}(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n})$$

$$G = \pi_0^* G' + \sum_{k=1}^{K} \pi_k^* \delta_{\theta_k^*}$$

$$(\pi_0^*, \pi_1^*, \ldots, \pi_K^*) \sim \mathrm{Dirichlet}(\alpha, n_1, \ldots, n_K)$$

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

- Use a stick-breaking representation for *G'* and truncate as before.
- No explicit ordering of the non-empty clusters makes for better mixing.
- Explicit representation of *G* allows for posterior estimates of functionals of *G*.



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

$$\theta_i|G \sim G$$

$$x_i|\theta_i \sim F(\theta_i)$$

Other Inference Algorithms

- Split-merge algorithms [Jain & Neal 2004].
  - Close in spirit to reversible-jump MCMC methods [Green & richardson 2001].

- Sequential Monte Carlo methods [Liu 1996, Ishwaran & James 2003, Fearnhead 2004, Mansingkha et al 2007].

- Variational algorithms [Blei & Jordan 2006, Kurihara et al 2007, Teh et al 2008].

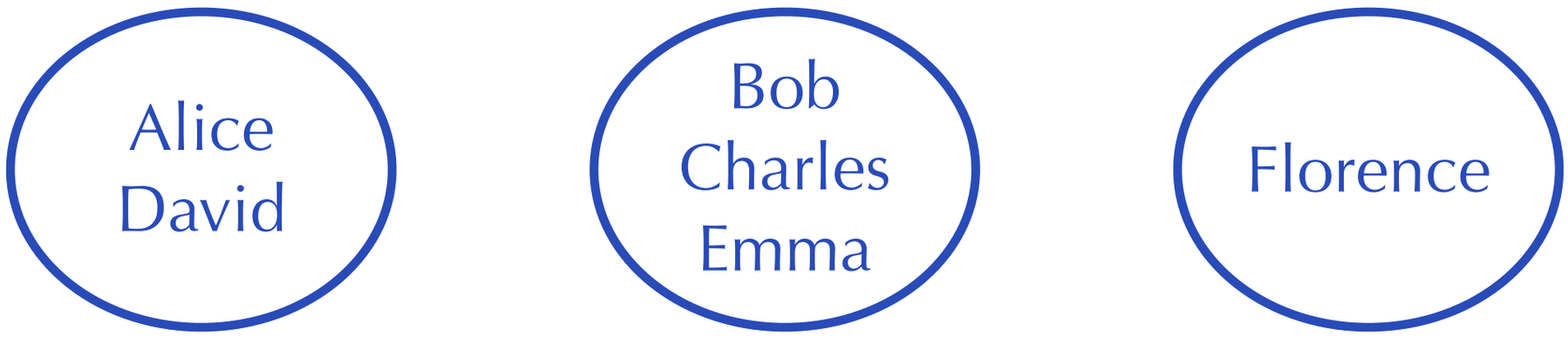
- Expectation propagation [Minka & Ghahramani 2003, Tarlow et al 2008].

Pitman-Yor Process

[Aldous 1985, Pitman 2006]

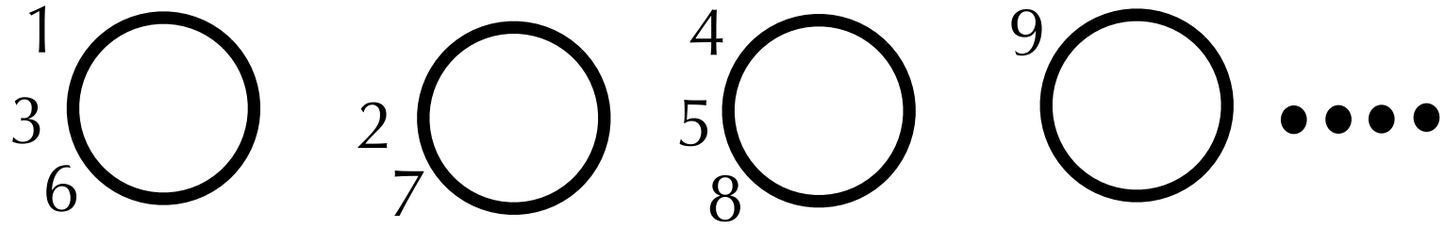
Partitions

- A **partition** ϱ of a set S is:
  - A disjoint family of non-empty subsets of S whose union in S.
  - S = {Alice, Bob, Charles, David, Emma, Florence}.
  - ϱ = { {Alice, David}, {Bob, Charles, Emma}, {Florence} }.



- Denote the set of all partitions of S as $\mathcal{P}_S$.
- **Random partitions** are random variables taking values in $\mathcal{P}_S$.
- We will work with partitions of $S = [n] = \{1,2,...n\}$.

Chinese Restaurant Process



- Each customer comes into restaurant and sits at a table:

$$p(\text{sit at table } c) = \frac{n_c}{\alpha + \sum_{c \in \varrho} n_c}$$

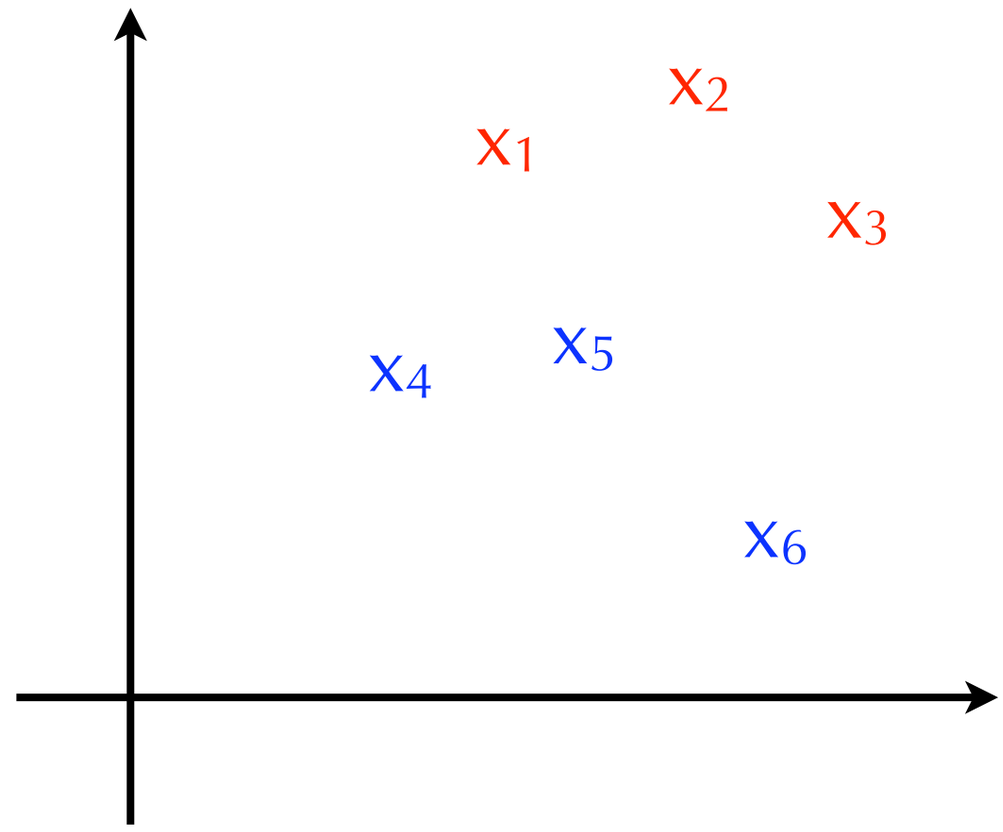
$$p(\text{sit at new table}) = \frac{\alpha}{\alpha + \sum_{c \in \varrho} n_c}$$

- Customers correspond to elements of set S, and tables to clusters in the partition ϱ.
- Multiplying conditional probabilities together, we get the overall probability of ϱ:

$$P(\varrho|\alpha) = \frac{\alpha^{|\varrho|}\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{c \in \varrho} \Gamma(|c|)$$

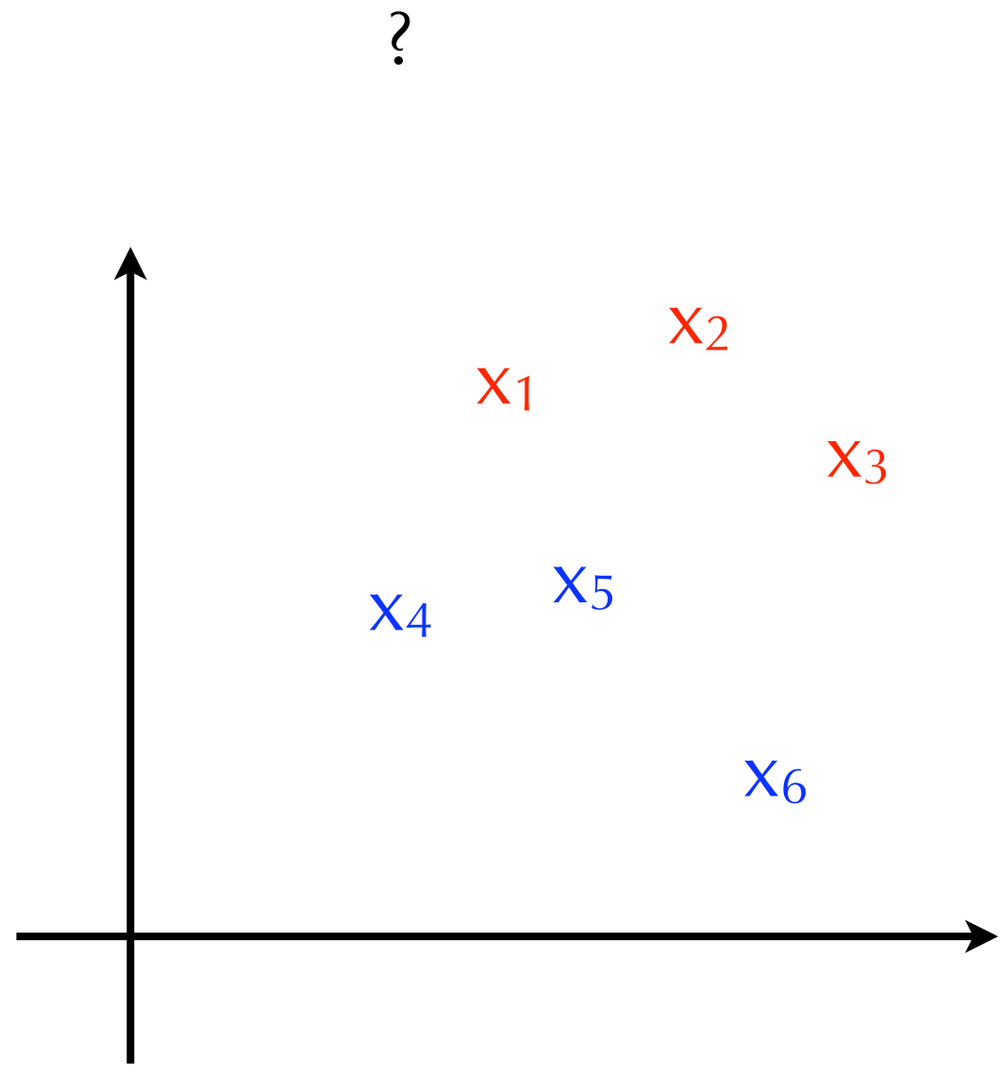
Projectivity and Exchangeability

Projective and Exchangeable Models of Data



Projective and Exchangeable Models of Data

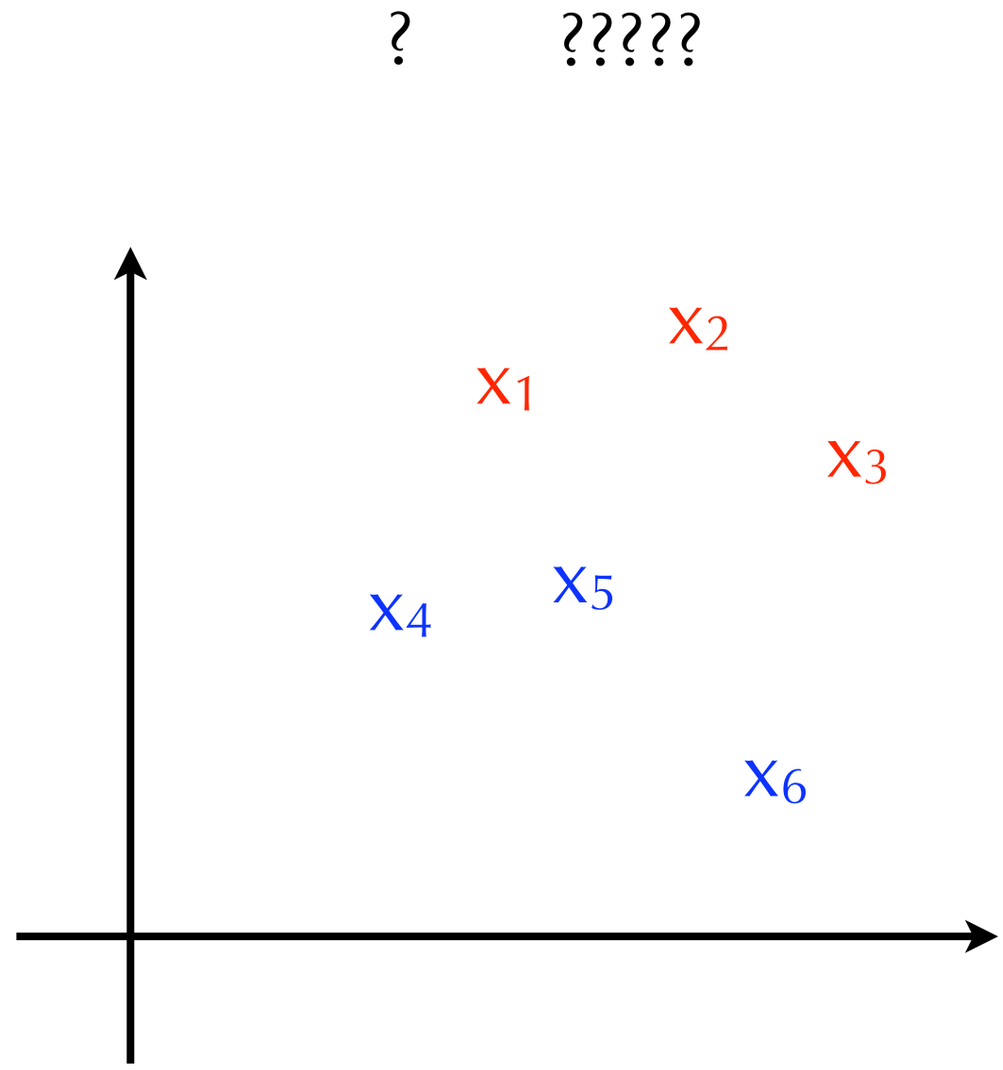
- There will be 1 test item.
  Will this change your predictions?



Projective and Exchangeable Models of Data

- There will be 1 test item.
  Will this change your predictions?

- There will be 5 additional test items.
  Will this change your predictions?

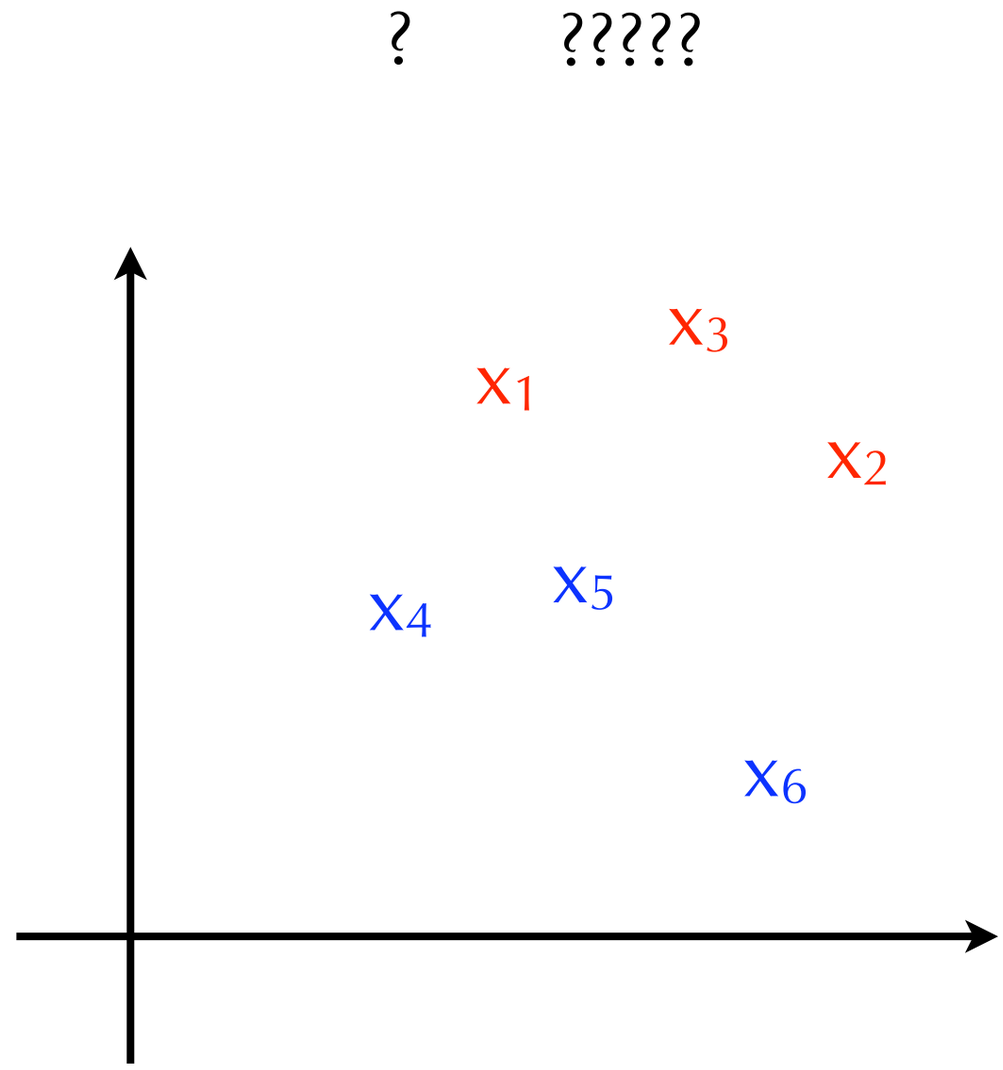


Projective and Exchangeable Models of Data

- There will be 1 test item.
  Will this change your predictions?

- There will be 5 additional test items.
  Will this change your predictions?

- Item labels were permuted.
  Will this change your predictions?



Consistency and Projectivity

- Let ϱ be a partition of S, and S' $\subset$ S be a subset. The **projection** of ϱ onto S' is the partition of S' defined by ϱ:

$$\text{PROJ}(\varrho, S') = \{\, c \cap S' \mid c \cap S' \neq \varnothing,\ c \in S \,\}$$

- I.e., all elements of *S* except those in *S'* are removed from ϱ.
- For example,

$$\text{PROJ}(\{\{1,3,6\},\{2,7\},\{4,5,8\},\{9\}\}, [6]) = \{\{1,3,6\},\{2\},\{4,5\}\}$$

Consistent/Projective Random Partitions

- A sequence of distributions $P_1, P_2, \ldots$ over $\mathcal{P}_{[1]}, \mathcal{P}_{[2]}, \ldots$ is **projective** or **consistent** if

$$\begin{aligned} \rho_m &\sim P_m \\ \rho_n &= \mathrm{PROJ}(\rho_m, [n]) \end{aligned} \quad \Rightarrow \quad \rho_n \sim P_n$$

$$P_m(\{\rho_m : \mathrm{PROJ}(\rho_m, [n]) = \rho_n\}) = P_n(\rho_n)$$

- Such a sequence can be extended to a distribution over $\mathcal{P}_{\mathbb{N}}$.
- The Chinese restaurant process is projective since:
  - The finite mixture model is, and
  - also it is defined sequentially.
- A projective model is one that does not change when more data items are introduced (and can be learned sequentially in a self-consistent manner).

Exchangeable Random Partitions

- A distribution over partitions $\mathcal{P}_S$ is **exchangeable** if it is invariant to permutations of S: For example,

$$P(\varrho = \{\{1,3,6\},\{2,7\},\{4,5,8\},\{9\}\}) =$$

$$P(\varrho = \{\{\sigma(1), \sigma(3), \sigma(6)\},\{\sigma(2), \sigma(7)\},\{\sigma(4), \sigma(5), \sigma(8)\},\{\sigma(9)\}\})$$

  where S = [9] = {1,...,9}, and σ is a permutation of [9].

- The Chinese restaurant process satisfies exchangeability:
  - The finite mixture model is exchangeable (iid given parameters).
  - The probability of ϱ under the CRP does not depend on the identities of elements of *S*.
- An exchangeable is one that does not depend on the (arbitrary) way data items are indexed.

Infinitely Exchangeable Random Variables

- Let $x_1, x_2, x_3, \ldots$ be an **infinitely exchangeable** sequence of random variables:

$$P(x_1, \ldots, x_n) = P(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

  for all n and permutations σ of $[n]$.

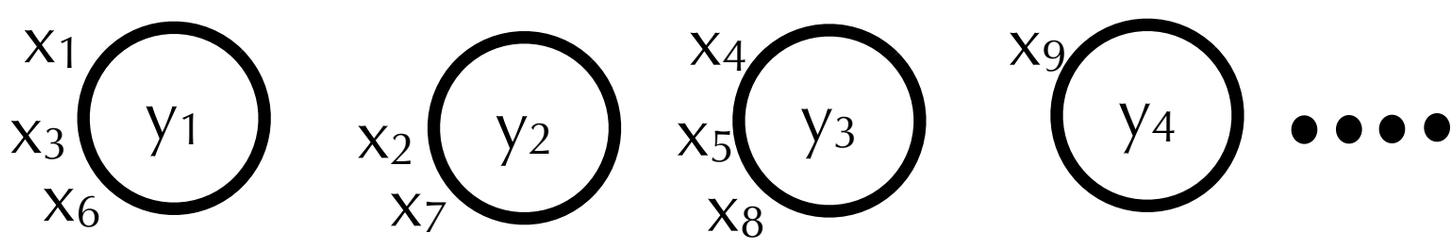
- Generalization of i.i.d. variables, and can be constructed as mixtures of such:

$$P(x_1, \ldots, x_n) = \int P(G) \prod_{i=1}^{n} P(x_i|G) dG$$

- **de Finetti's Theorem**: infinitely exchangeable sequences can always be represented as mixtures of i.i.d. variables. Further the latent parameter G is unique, called the **de Finetti measure**.

Dirichlet Process

- Since the CRP is projective and exchangeable, we can define an infinitely exchangeable sequence as follows:
  - Sample $\varrho \sim \mathrm{CRP}(\mathbb{N}, \alpha)$.
  - For $c \in \varrho$:
    - sample $y_c \sim H$.
  - For $i = 1,2,\ldots$:
    - set $x_i = y_c$ where $i \in c$.



- The resulting de Finetti measure is the DP with parameters α and H.

[Ferguson 1973, Blackwell & MacQueen 1973]

Why Infinitely Exchangeable Models?

- A model for a dataset $x_1,x_2,\ldots,x_n$ is a joint distribution $P(x_1,x_2,\ldots,x_n)$.
- An infinitely exchangeable model means:
  - The way data items are ordered or indexed does not matter.
  - Model is unaffected by existence of additional unobserved data items, e.g. test items.
    - To predict m additional test items, we would need

$$P(x_1,\ldots,x_n,\ x_{n+1},\ldots,x_{n+m})$$

    - If model is not infinitely exchangeable, predictive probabilities will be different for different values of m.
- There are scenarios where infinite exchangeability is suitable or unsuitable.

Exchangeability in Bayesian Statistics

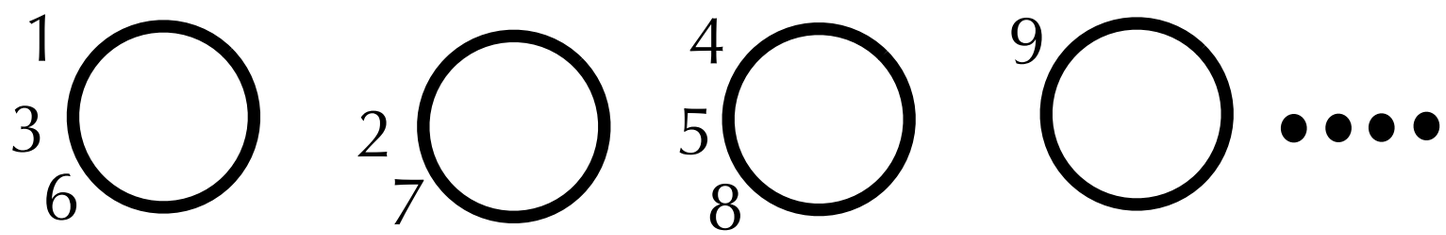
- Fundamental role of de Finetti's Theorem in Bayesian statistics:
  - From an assumption of exchangeability, we get a representation as a Bayesian model with a prior over the latent parameter.

$$P(x_1, \ldots, x_n) = \int P(G) \prod_{i=1}^{n} P(x_i|G) dG$$

- Generalizing infinitely exchangeable sequences lead to Bayesian models for richly structured data. E.g.,
  - exchangeability in network and relational data.
  - hierarchical exchangeability in hierarchical Bayesian models.
  - Markov exchangeability in sequence data.

Two-parameter Chinese Restaurant Process

- The **two-parameter Chinese restaurant process** CRP([n],d,α) is a distribution over $\mathcal{P}_{[n]}$: ($0 \leq d < 1$, $\alpha > -d$), described by the following process:



$$P(\text{sit at table } c) = \frac{n_c - d}{\alpha + \sum_{c \in \varrho} n_c} \qquad P(\text{sit at new table}) = \frac{\alpha + d|\varrho|}{\alpha + \sum_{c \in \varrho} n_c}$$

- Difference: **discount parameter** d.
  - Expect to get more tables, and more tables with few customers.

Pitman-Yor Process

- The EPPF under CRP([n],d,α) is:

$$P(\varrho) = \frac{[\alpha + d]_d^{|\varrho|-1}}{[\alpha + 1]_1^{n-1}} \prod_{c \in \varrho} [1 - d]_1^{|c|-1} \qquad [z]_b^m = z(z + b) \cdots (z + (m - 1)b)$$

- The two-parameter CRP is projective and exchangeable.

- The de Finetti measure is the **Pitman-Yor process**, which is a generalization of the Dirichlet process.

[Perman et al 1992, Pitman & Yor 1997, Ishwaran & James 2001]

Power-law Properties

Power-laws in Pitman-Yor Processes

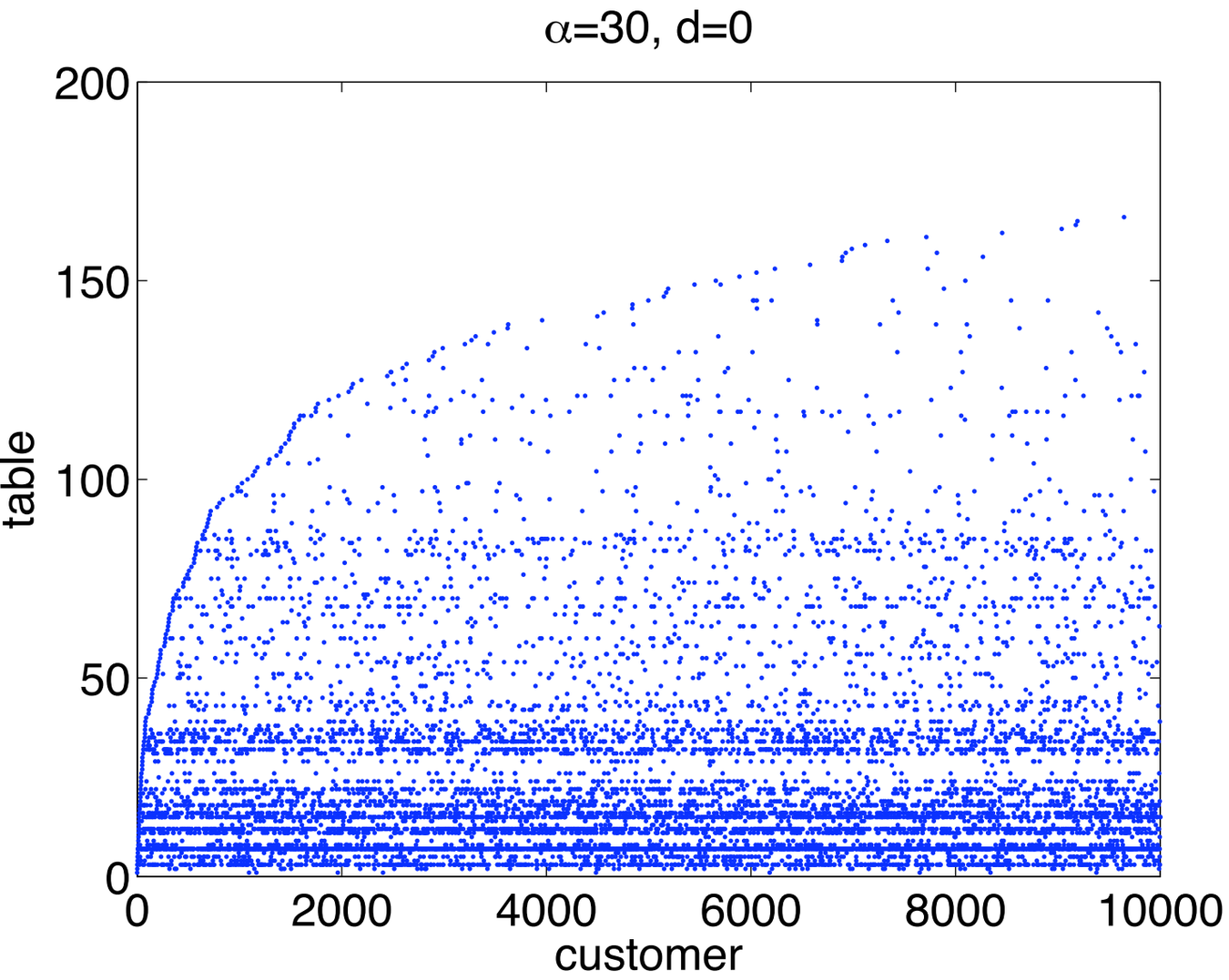
- Power-laws are commonly observed in nature and in human generated data.
- Pitman-Yor processes exhibit power-law properties and can be used to model data with such properties.

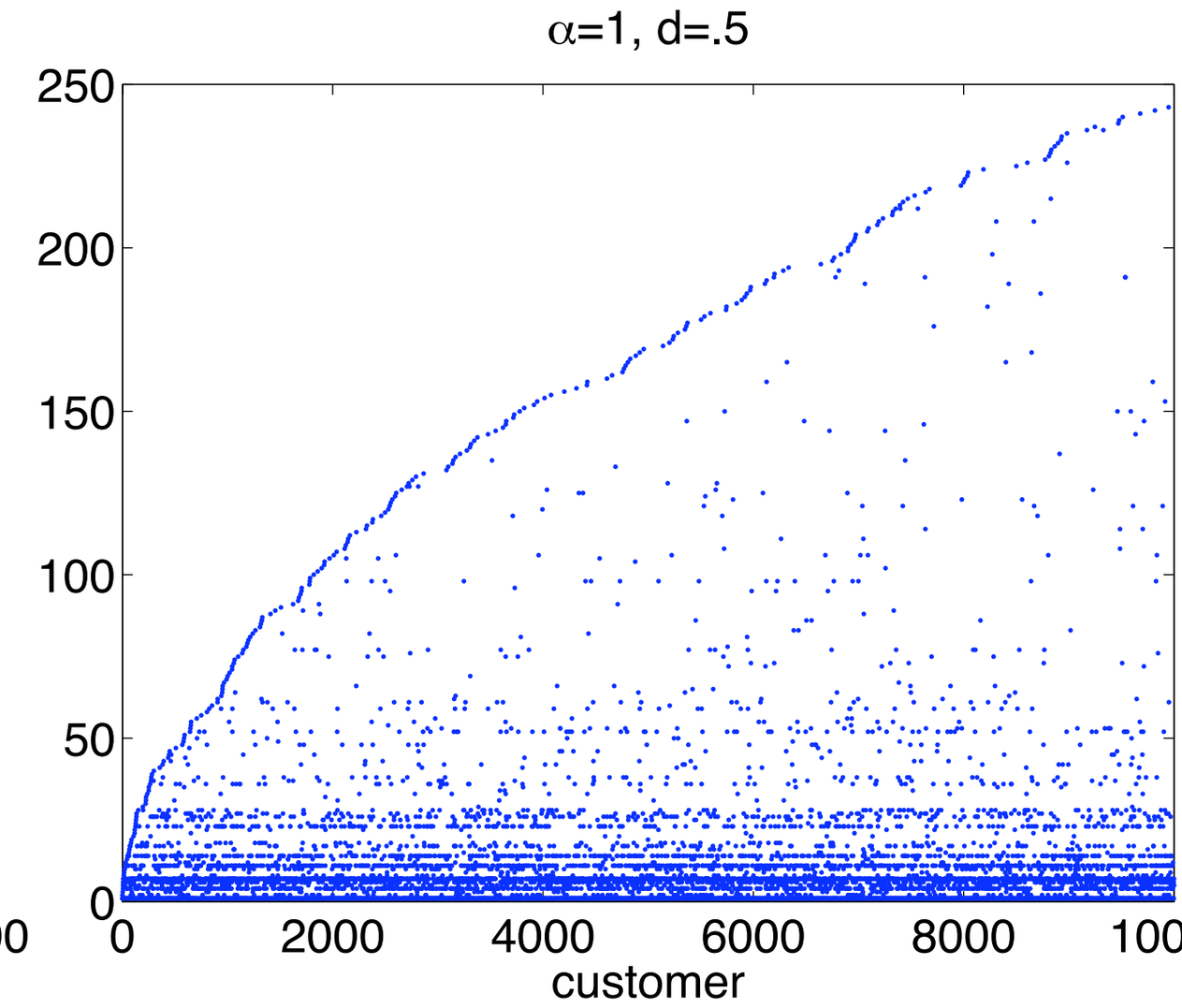
$$P(\text{sit at table } c) = \frac{n_c - d}{\alpha + \sum_{c \in \varrho} n_c} \qquad P(\text{sit at new table}) = \frac{\alpha + d|\varrho|}{\alpha + \sum_{c \in \varrho} n_c}$$

- With more occupied tables, chance of even more tables becomes higher.
- Tables with small occupancy numbers tend to have lower chance of getting new customers.

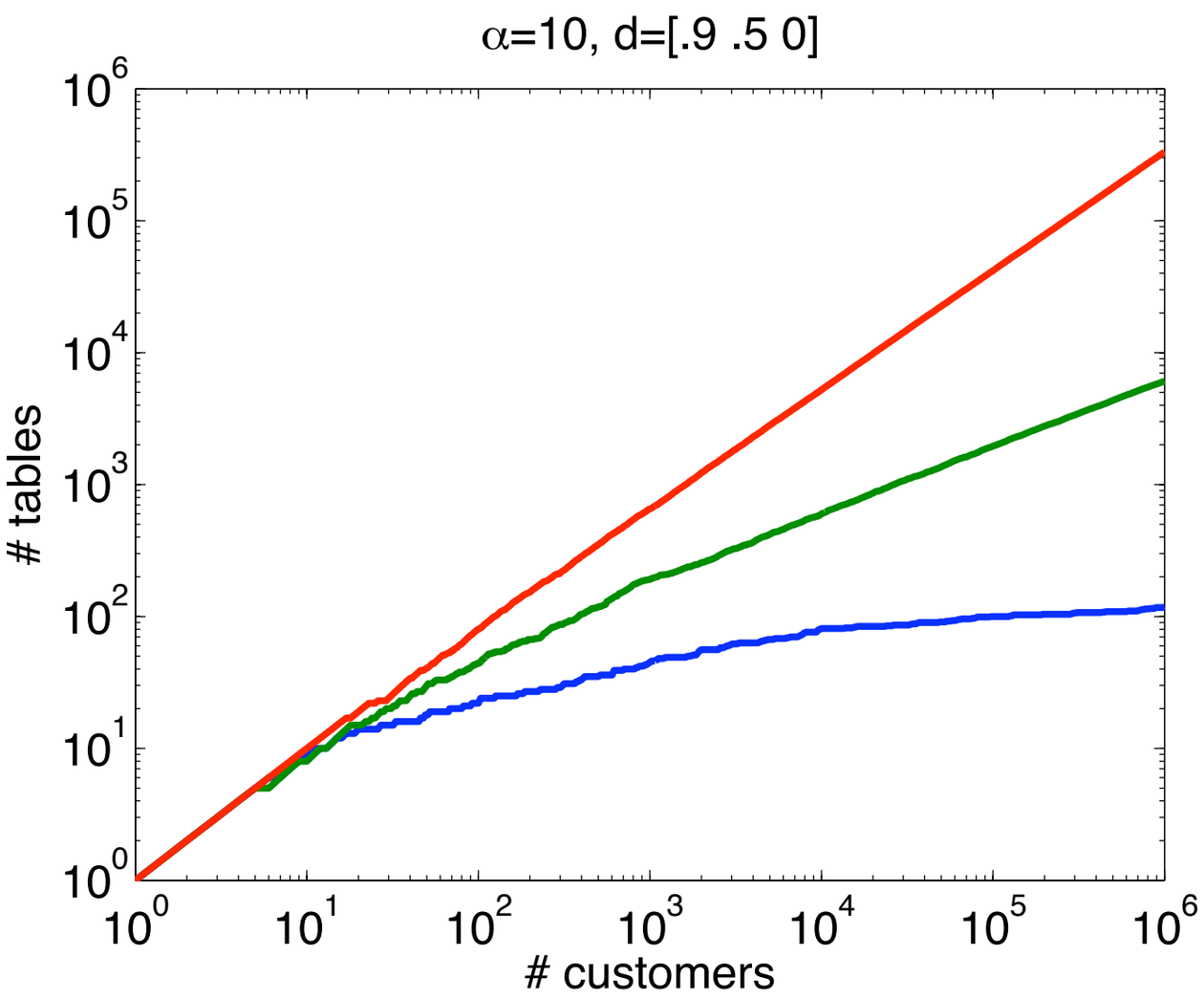
[Pitman 2006, Goldwater et al 2006, Teh 2006]

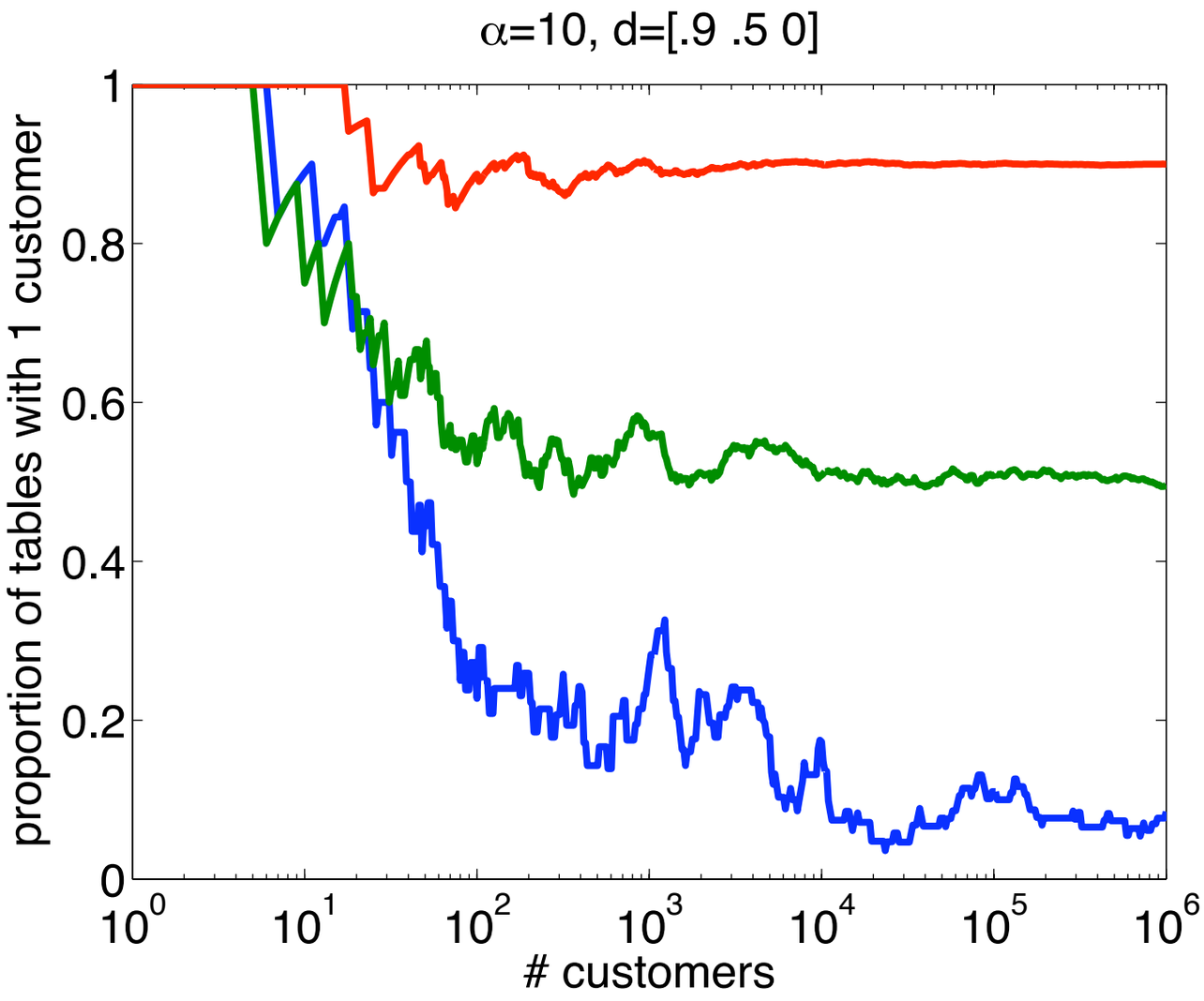
Power-Laws in Pitman-Yor Processes



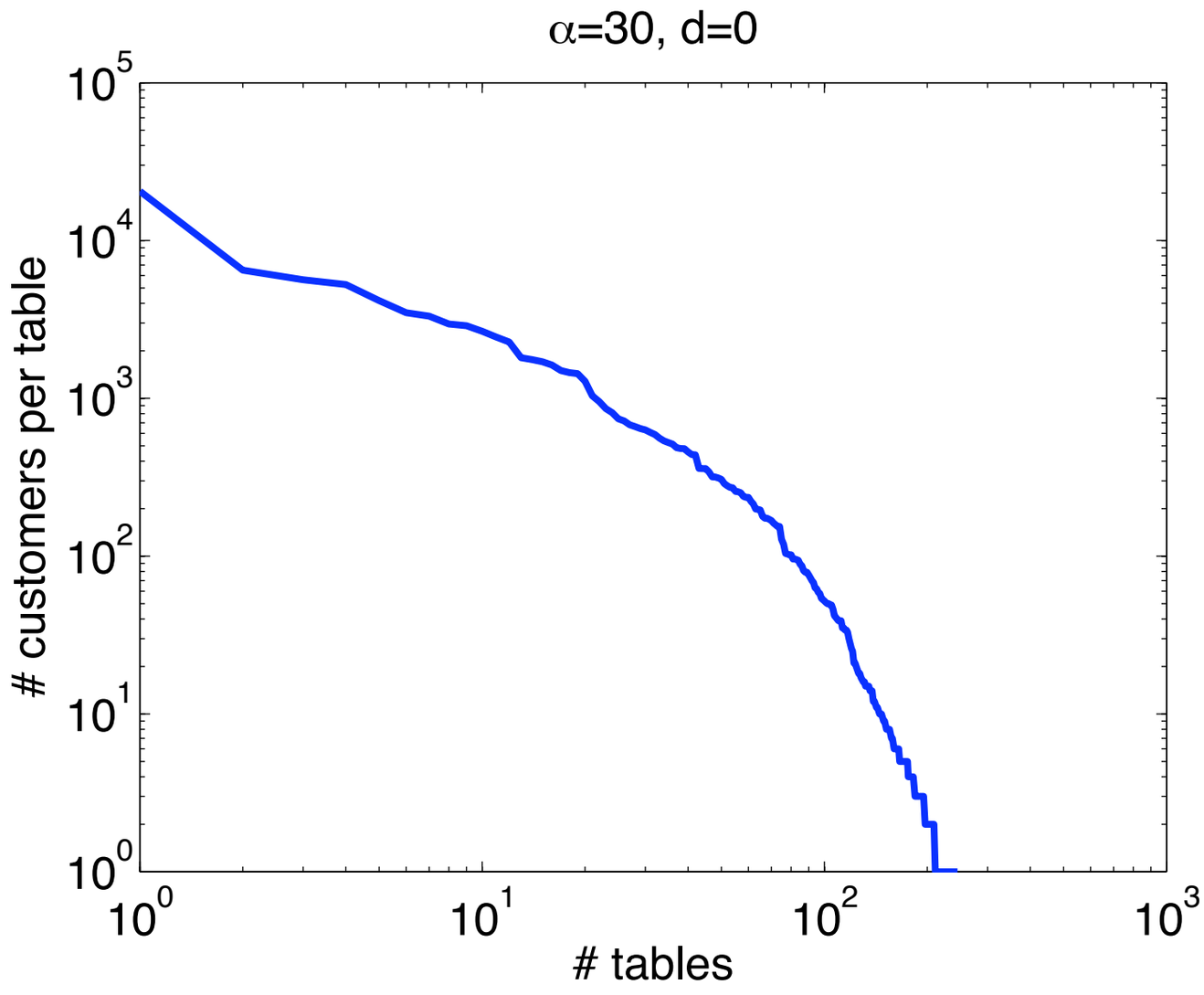


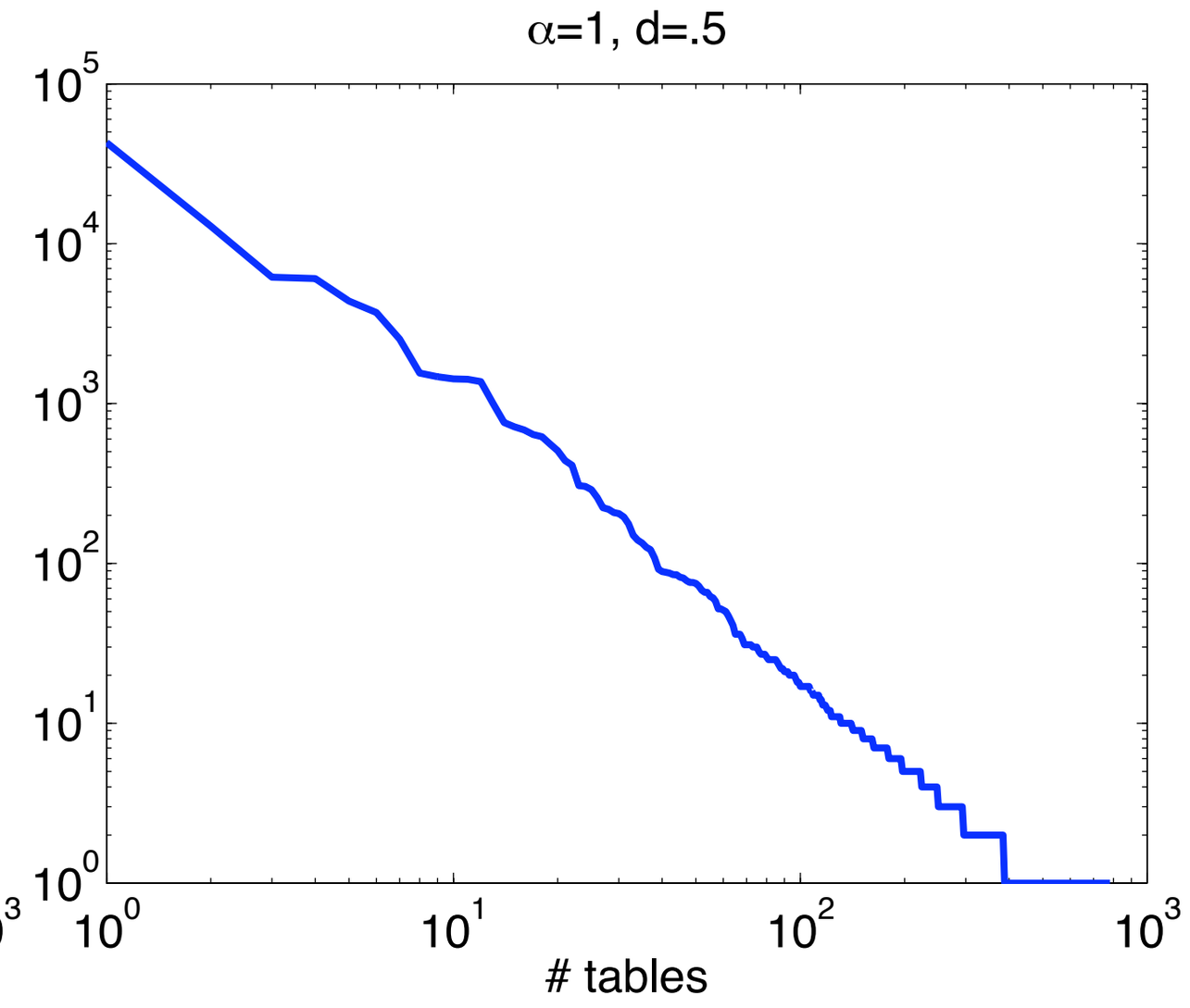
Power-Laws in Pitman-Yor Processes



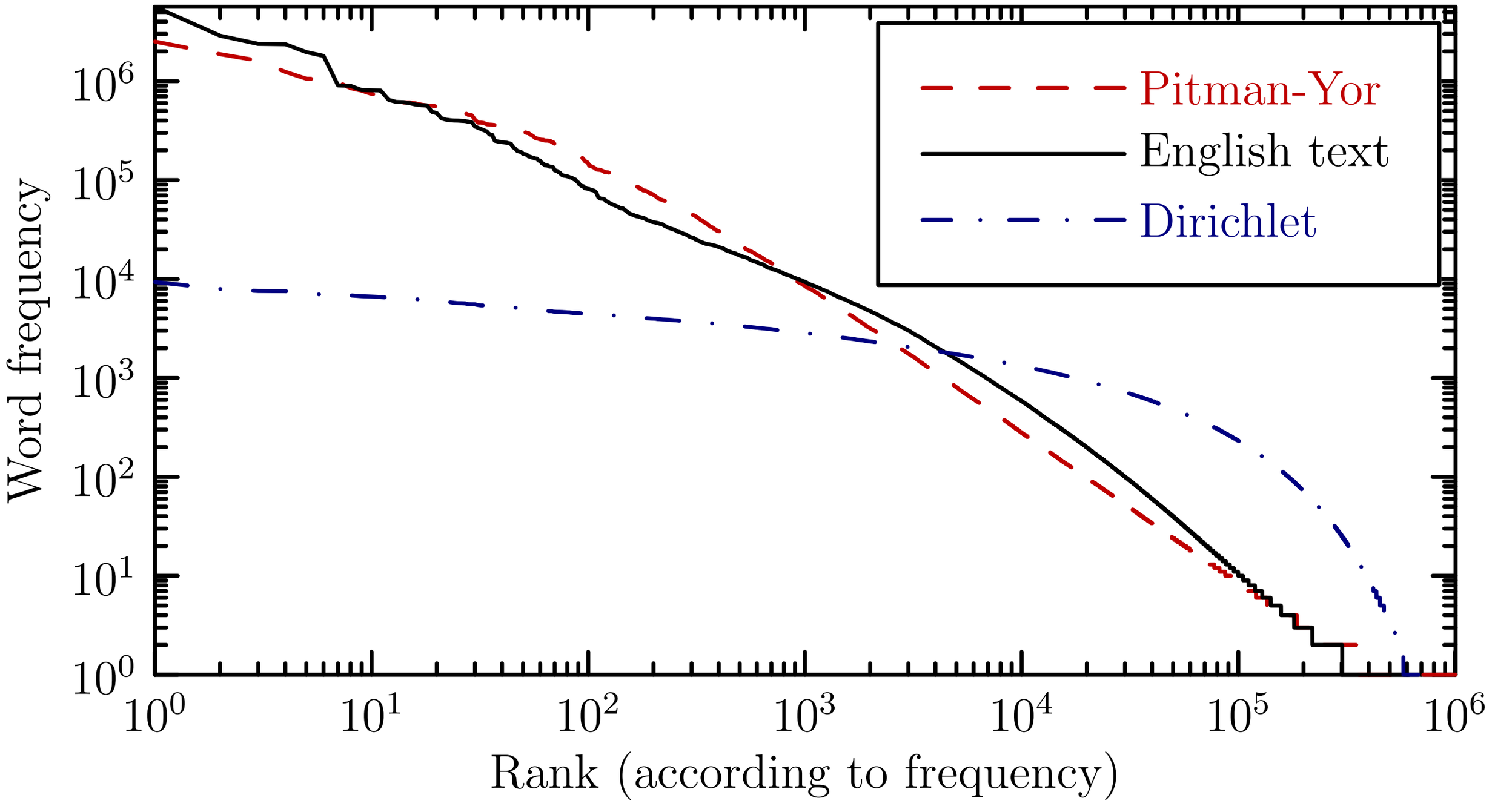


Power-Laws in Pitman-Yor Processes



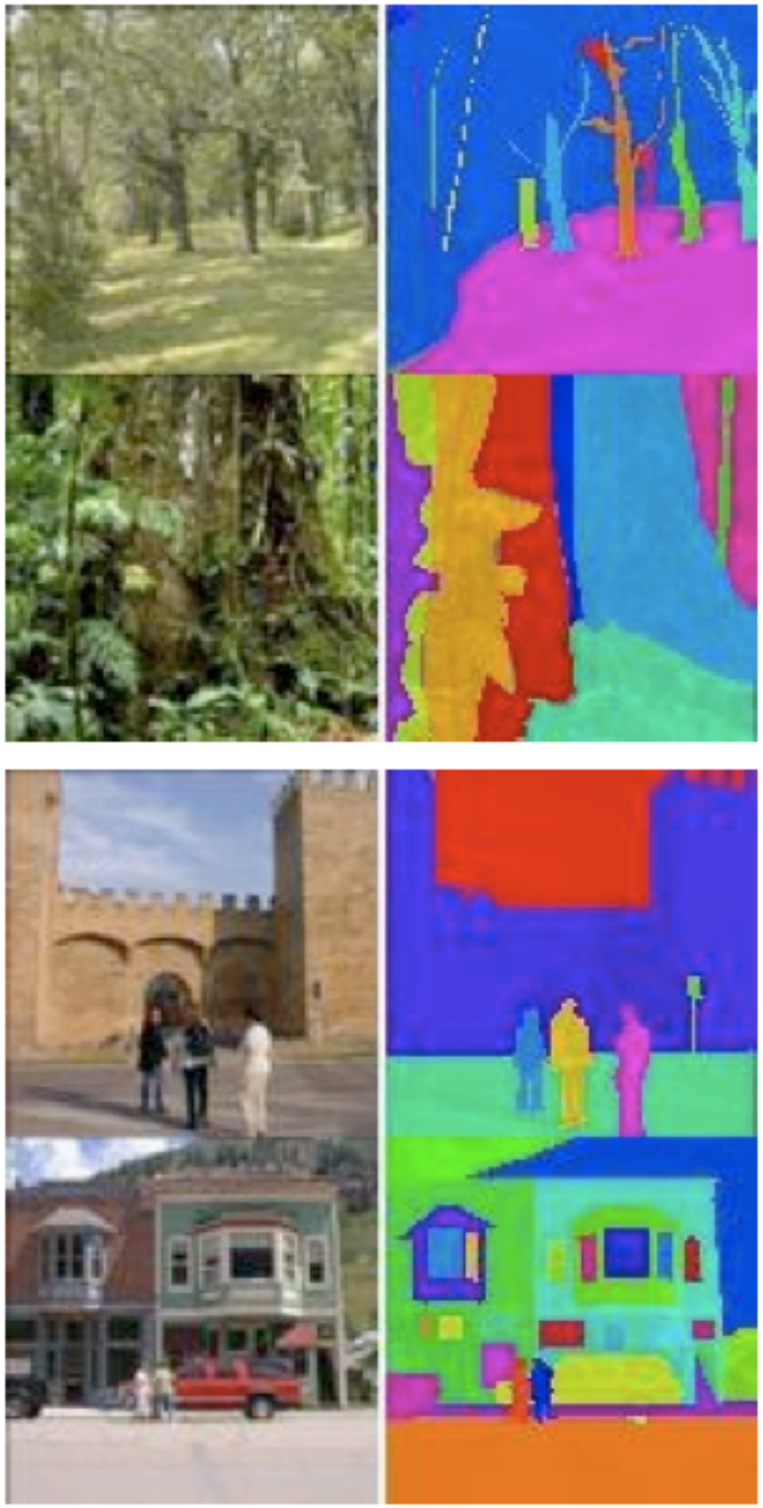


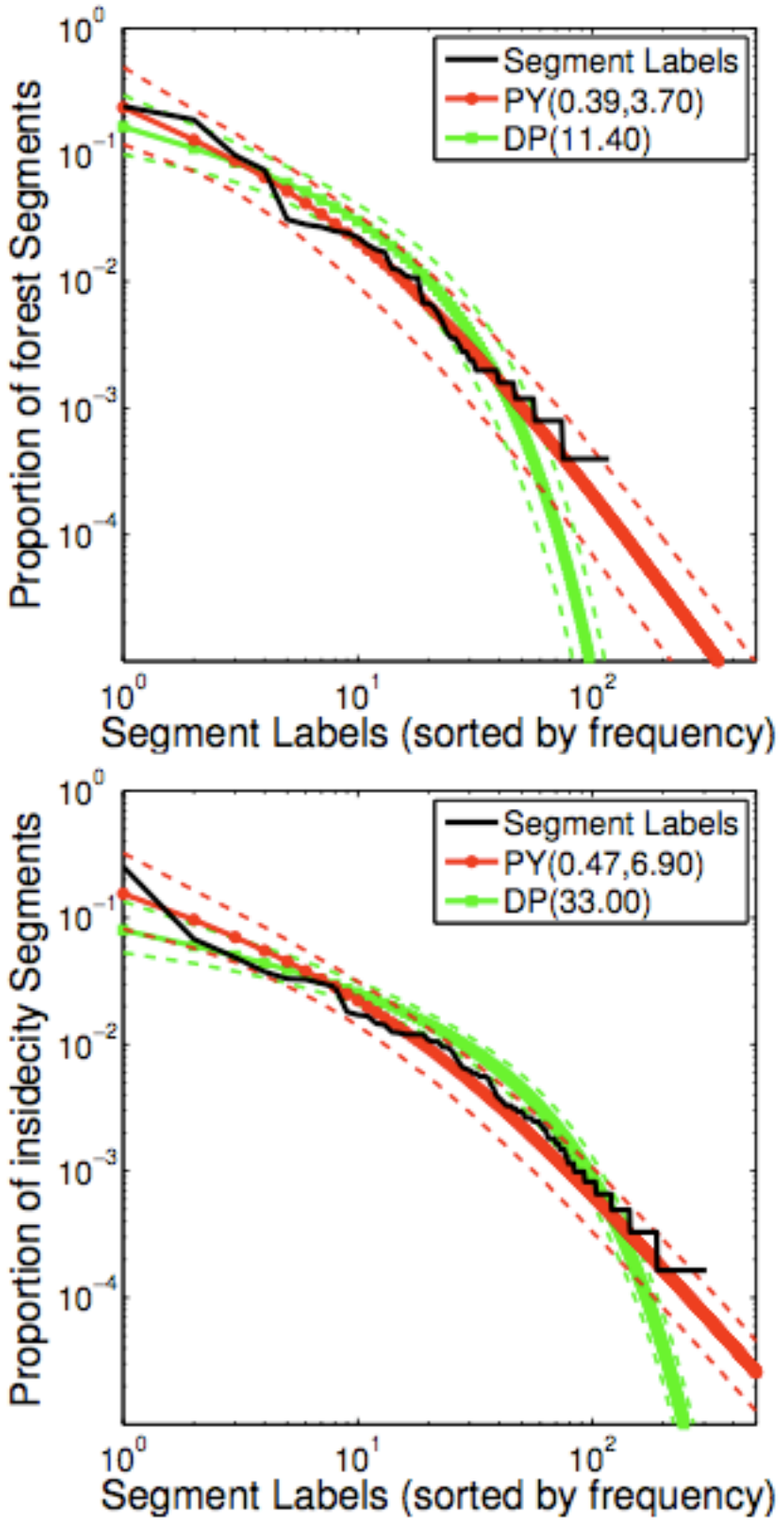
Power-law of English Word Frequencies

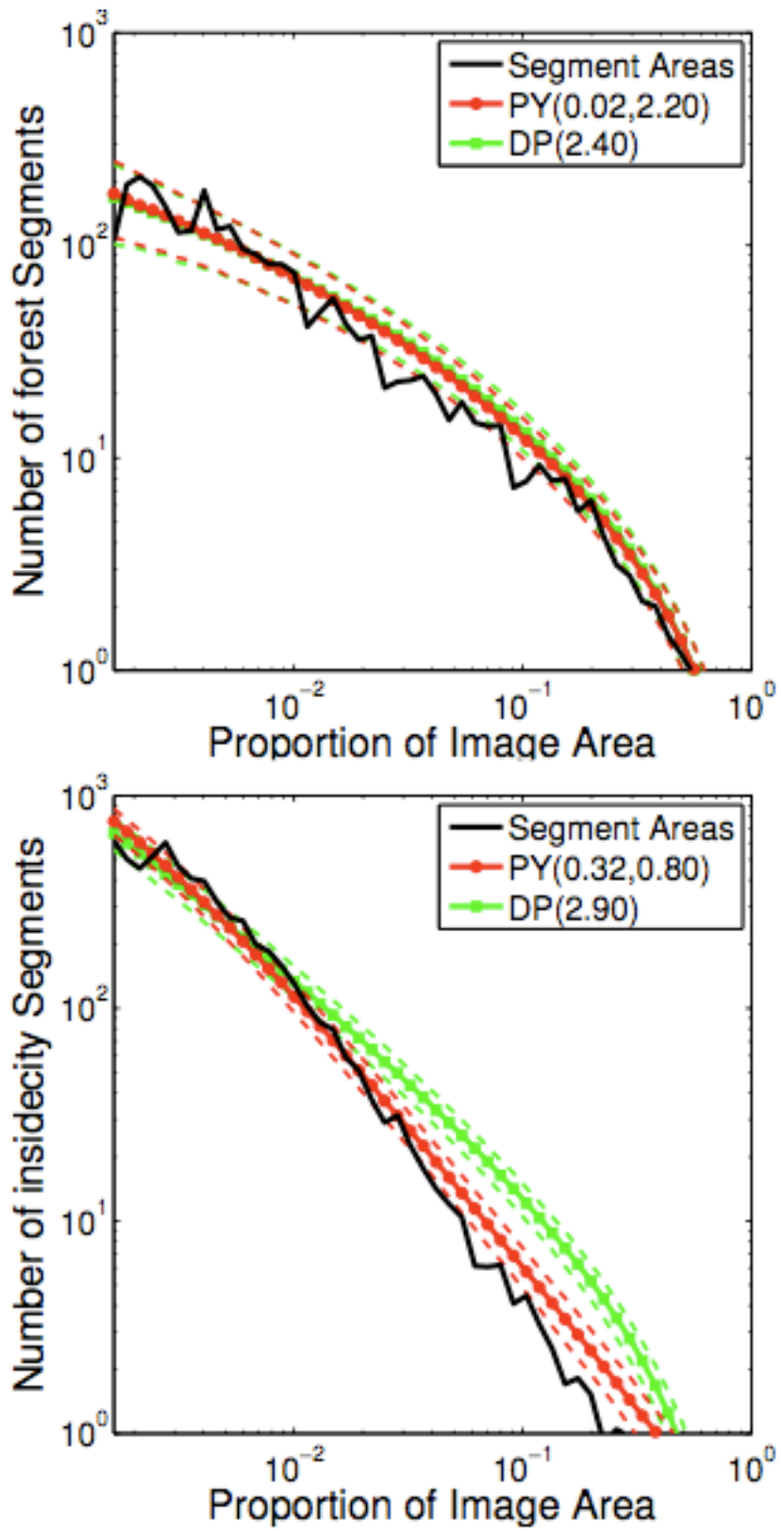


[Wood et al 2011]

Power-law of Image Segmentations







[Sudderth & Jordan 2009]

Pitman-Yor Process

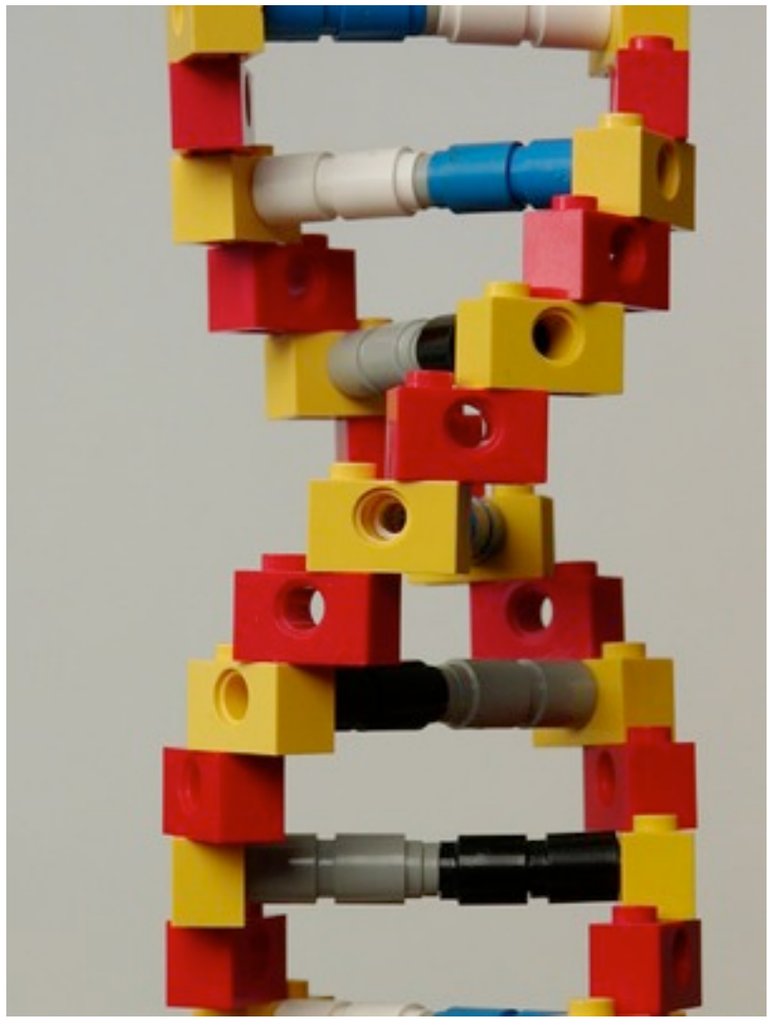
- Pitman-Yor processes have been applied in domains with power-laws:
  - computational linguistics;
  - computer vision.
- They also have interesting properties related to fragmentations and coagulations of partitions which can be used to build effective methods for sequence modelling and text compression.
- They also have stick-breaking constructions and are the next simplest generalization of Dirichlet processes. Other generalizations include:
  - normalized random measures;
  - species sampling models;
  - stick-breaking processes.

Hierarchical Bayesian Nonparametric Models

[Teh & Jordan 2010]

Nonparametric Building Blocks

- Easy to construct complex probabilistic models from simpler parts.
- Nonparametric Bayesian models are new classes of components for the statistical modeller.
  - Dependent random measures;
  - Hierarchical nonparametric models.
  - Nested models.



Hierarchical Dirichlet Process

[Teh et al 2006]

Topic Modelling

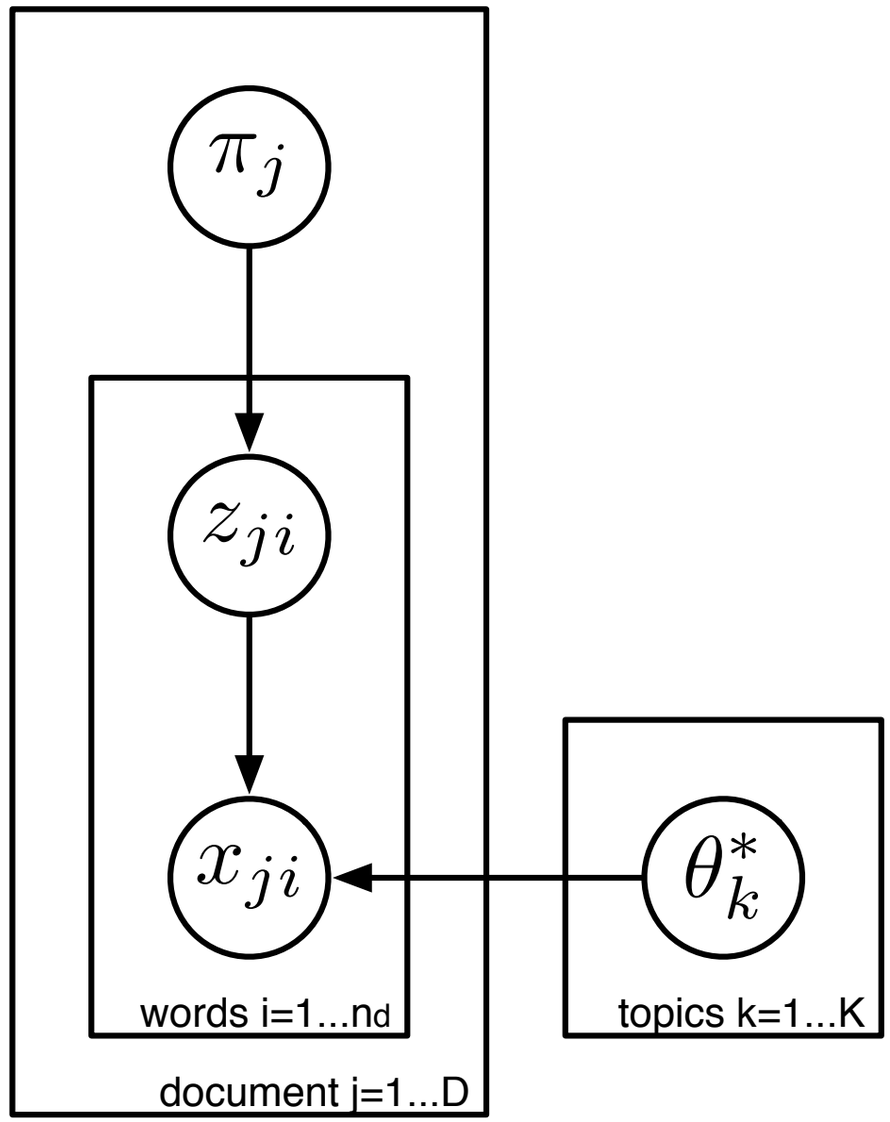
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

[Blei et al 2003, Griffiths & Steyvers 2004]

Latent Dirichlet Allocation

- Model a topic as a distribution over words that tend to co-occur together among documents.
- Model words in documents as exchangeable and documents as mixtures of topics.

$$\begin{aligned}
\pi_j &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
\theta_k^* &\sim \text{Dirichlet}(\beta/W, \ldots, \beta/W) \\
z_{ji}|\pi_j &\sim \text{Discrete}(\pi_j) \\
x_{ji}|z_{ji}, \theta_{z_{ji}}^* &\sim \text{Discrete}(\theta_{z_{ji}}^*)
\end{aligned}$$

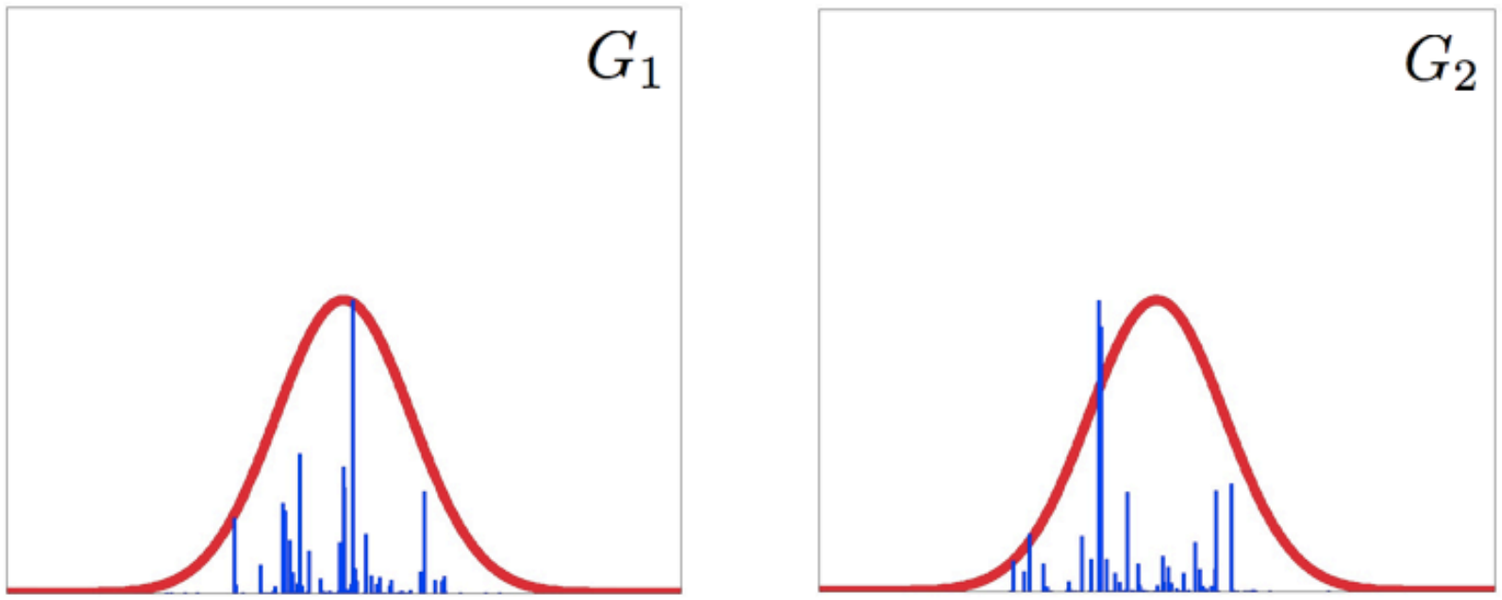


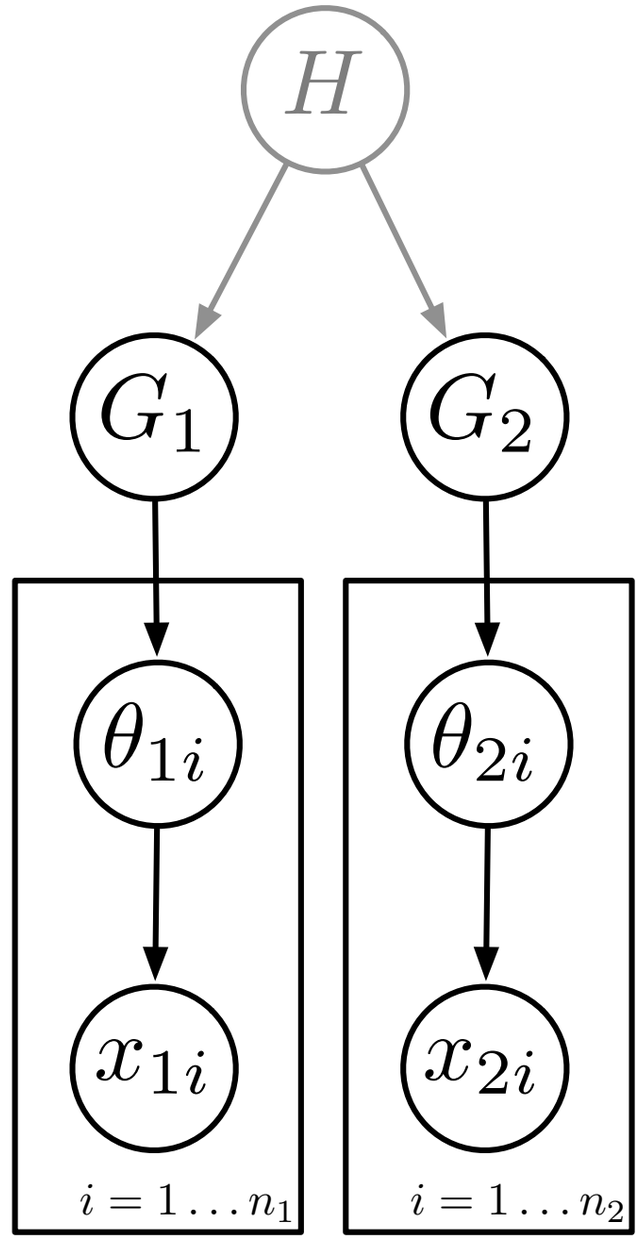
- How many topics can we find in a corpus?

[Blei et al 2003, Griffiths & Steyvers 2004]

Nonparametric Latent Dirichlet Allocation?

- Use a DP for each document.





- There is no sharing of topics across different documents, because *H* is smooth.
- Solution: make *H* discrete.
- Put a DP prior on *H*.

Hierarchical Dirichlet Process

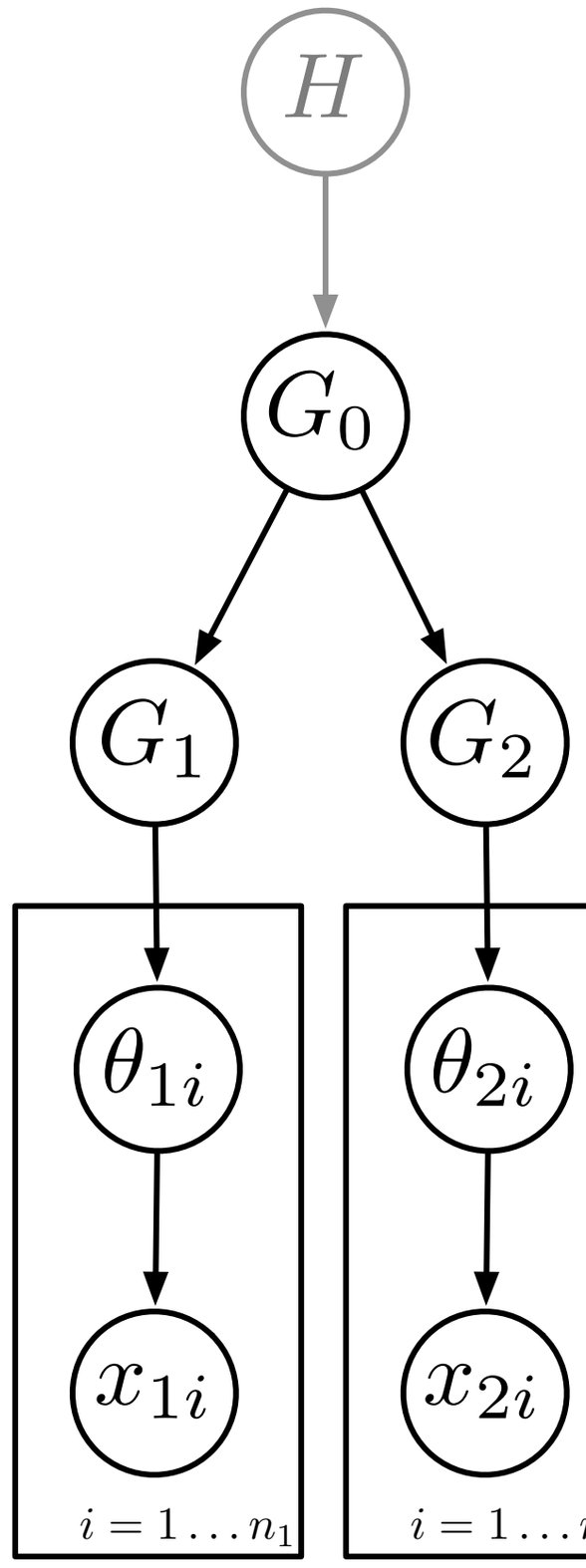
- A hierarchy of Dirichlet processes:

$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_1|G_0 \sim \mathrm{DP}(\alpha_1, G_0)$$
$$G_2|G_0 \sim \mathrm{DP}(\alpha_2, G_0)$$

- Extension to larger hierarchies straightforward:

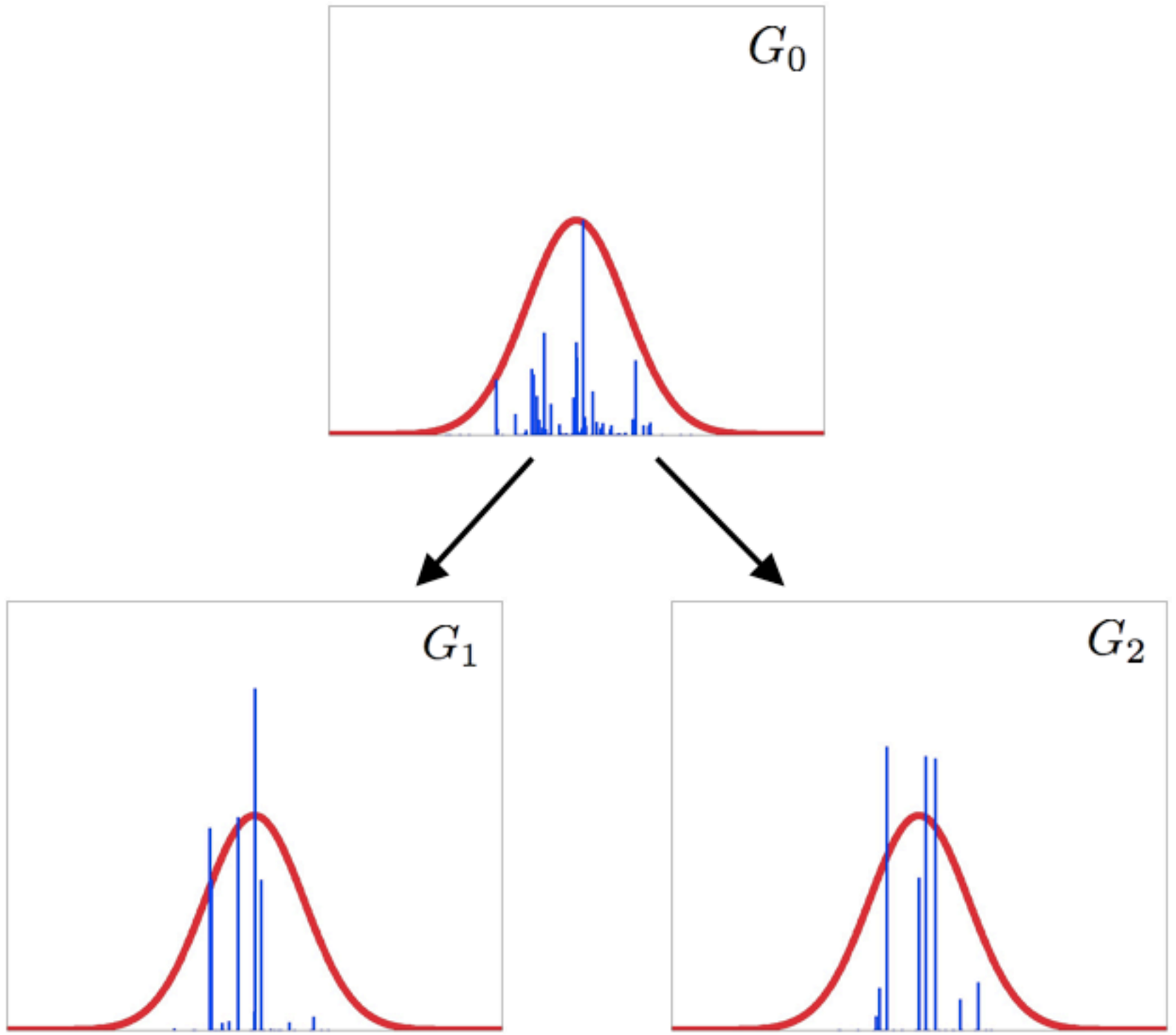
$$G_j \sim \mathrm{DP}(\alpha_j, G_{\mathrm{pa}(j)})$$

- Hierarchical modelling are a widespread technique to share statistical strength.

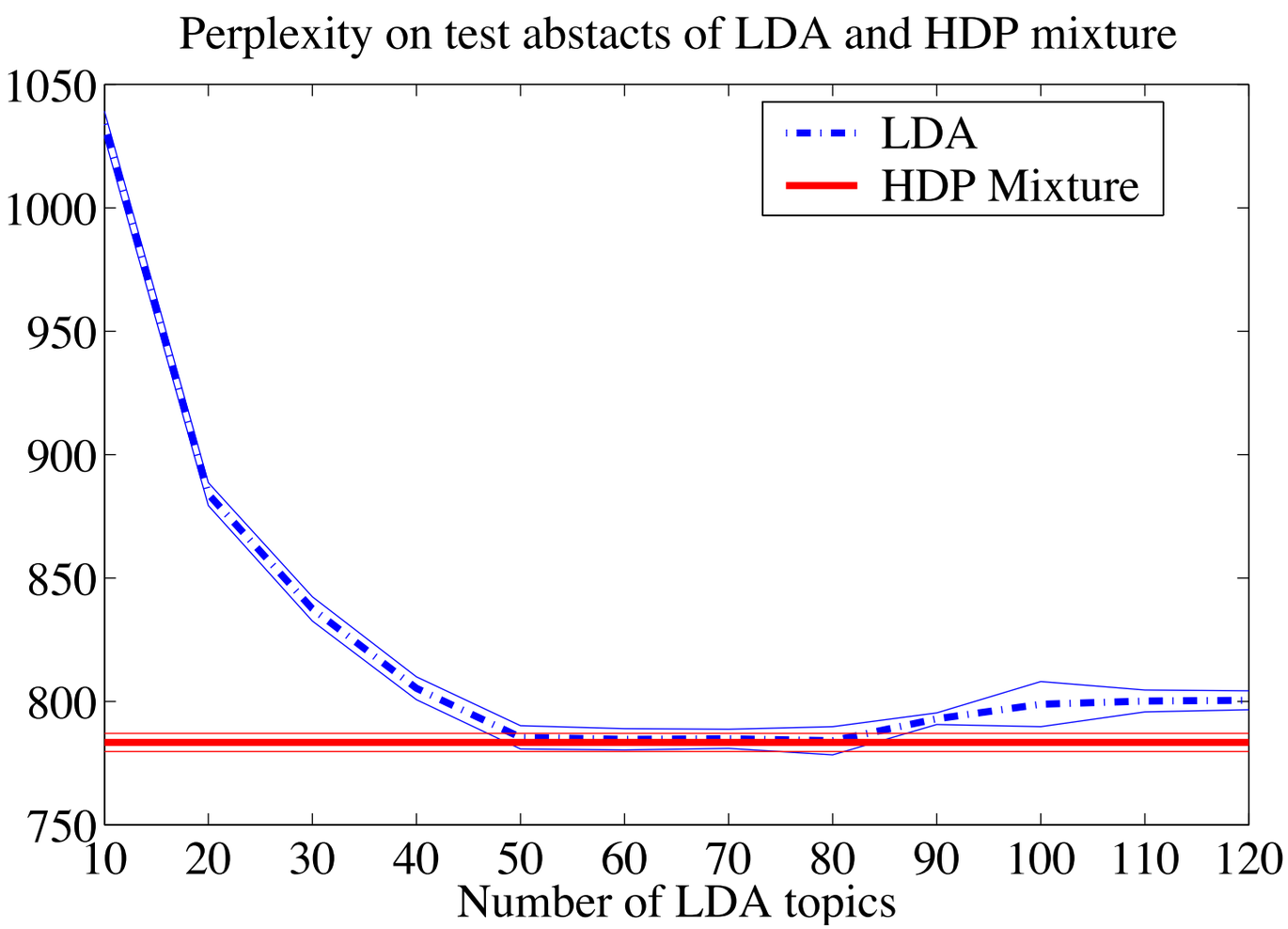


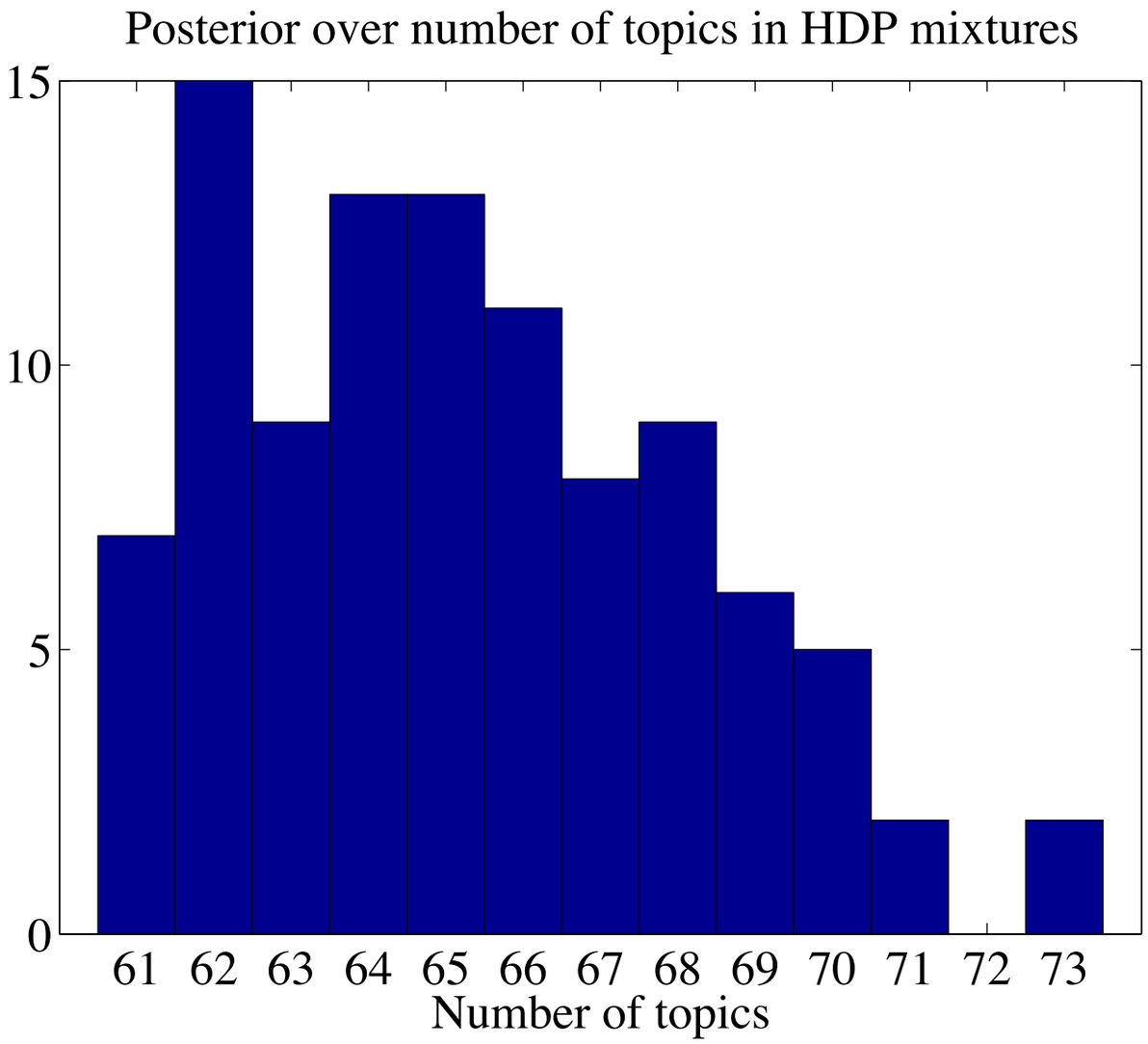
[Teh et al 2006]

Hierarchical Dirichlet Process

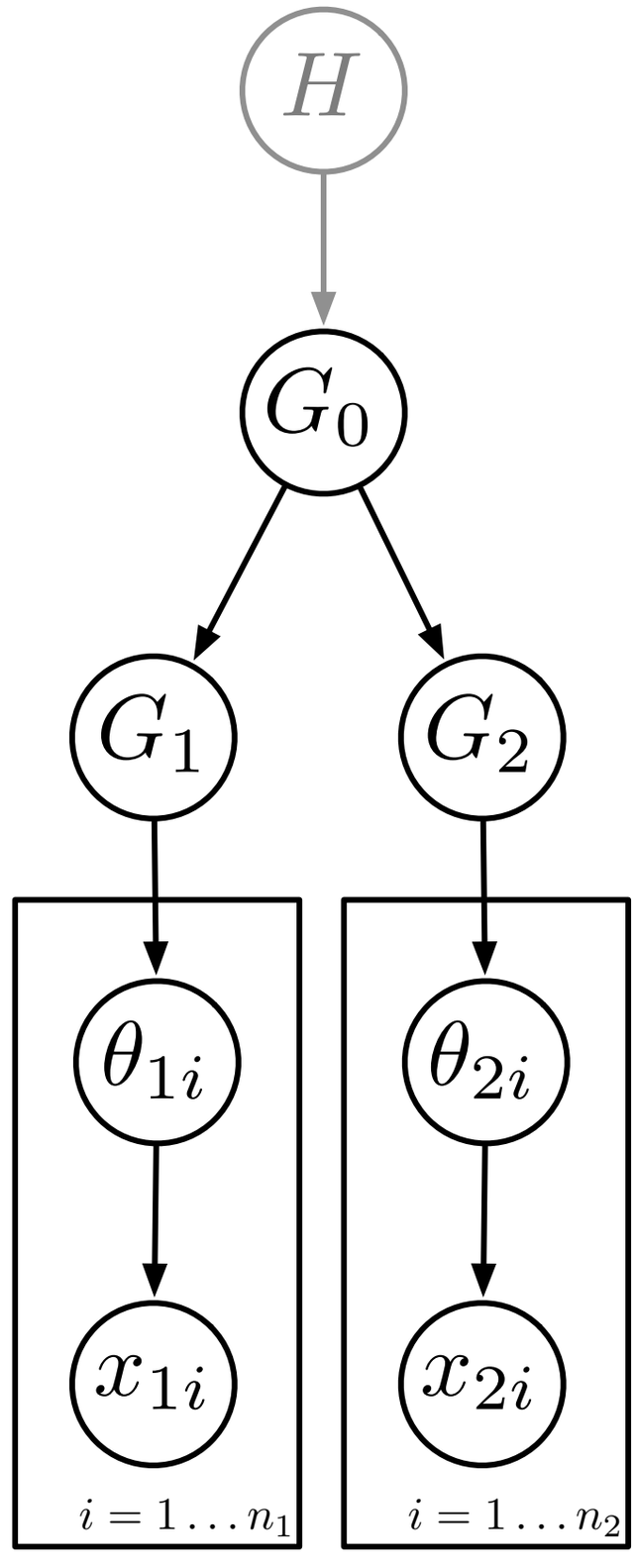


HDP-LDA


Perplexity on test abstacts of LDA and HDP mixture
LDA
HDP Mixture
1050
1000
950
900
850
800
750
10
20
30
40
50
60
70
80
90
100
110
120
Number of LDA topics

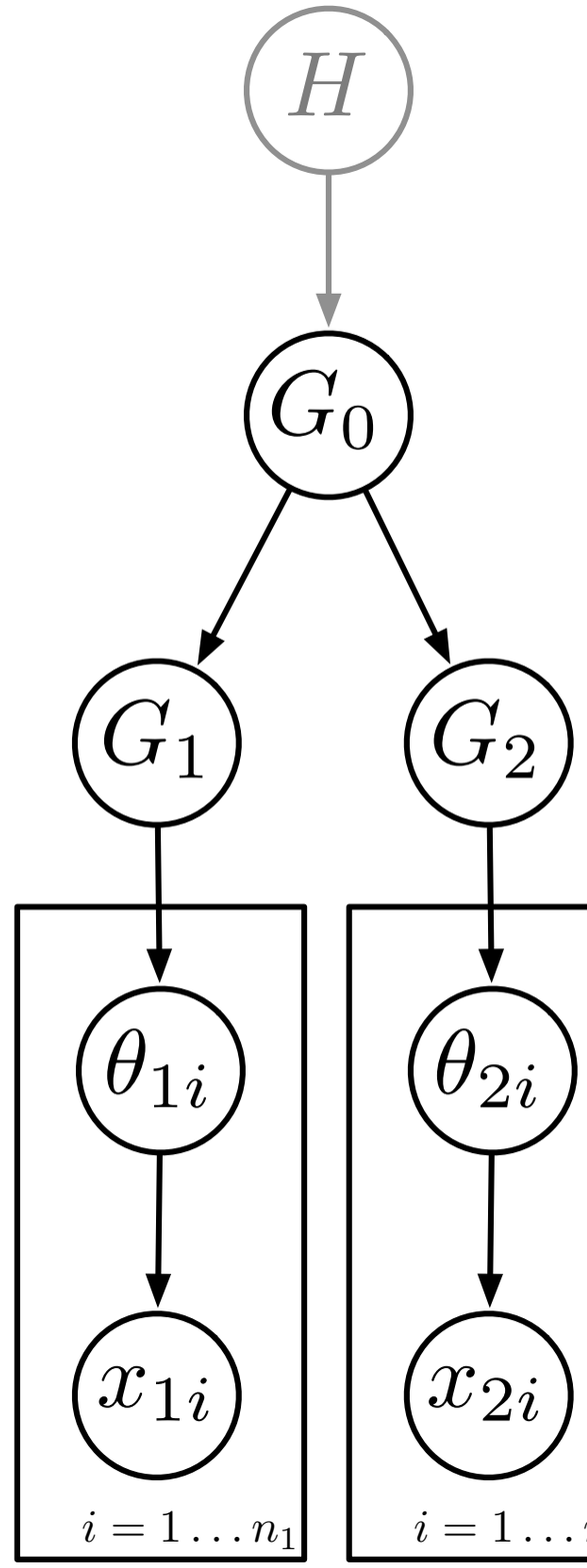

Posterior over number of topics in HDP mixtures
15
10
5
0
61
62
63
64
65
66
67
68
69
70
71
72
73
Number of topics

Chinese Restaurant Franchise



- G_1 and G_2 can both be represented using CRPs.

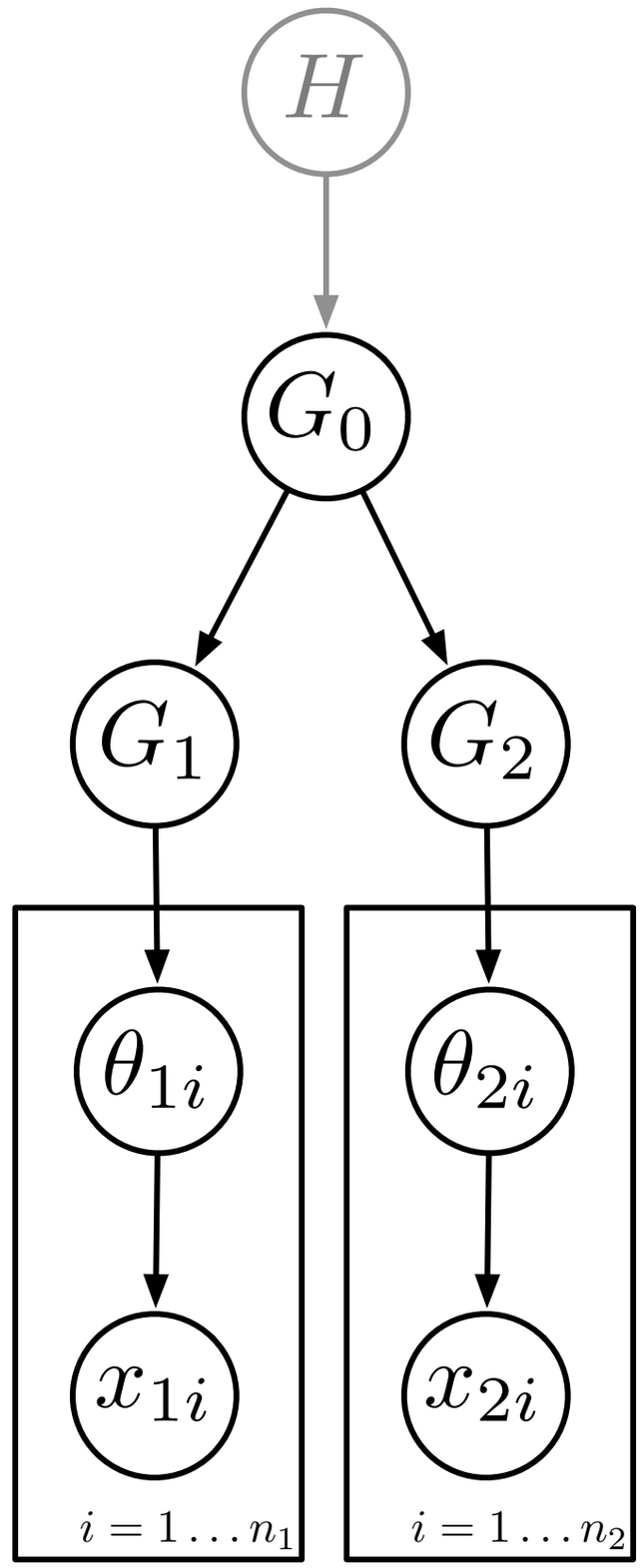
Chinese Restaurant Franchise

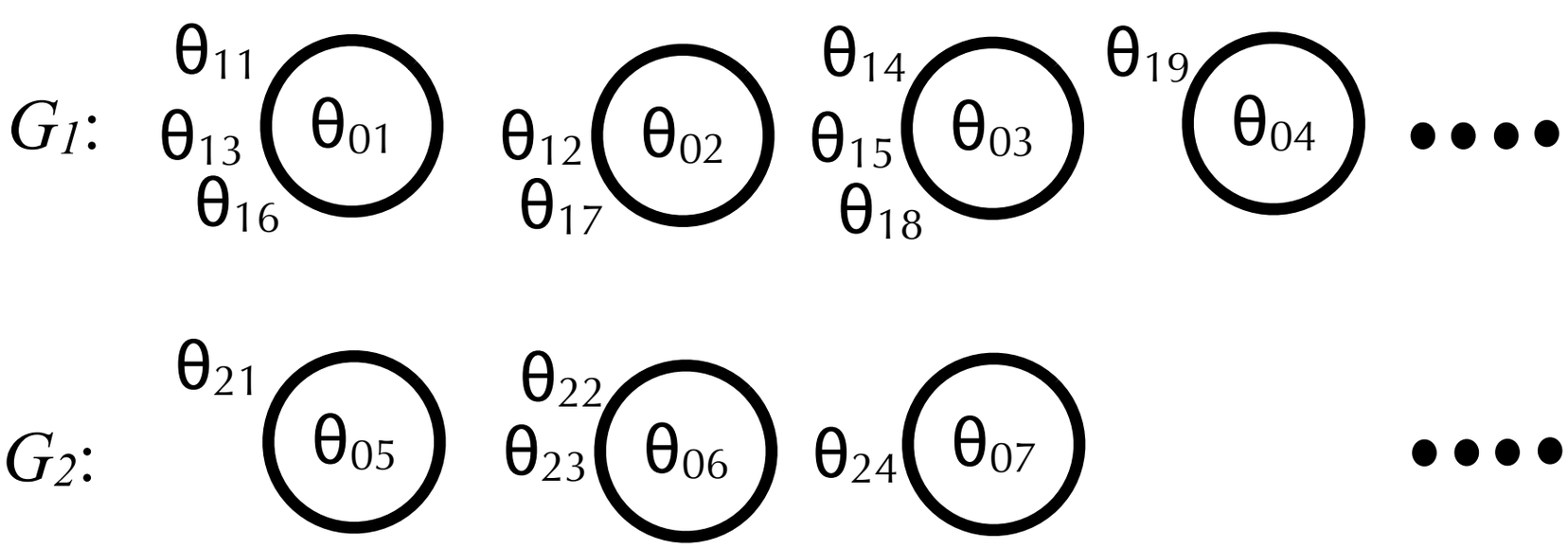


G_1: θ_{11} θ_{13} θ_{16} θ_{01} θ_{12} θ_{17} θ_{02} θ_{14} θ_{15} θ_{18} θ_{03} θ_{19} θ_{04} ••••

- G_1 and G_2 can both be represented using CRPs.

Chinese Restaurant Franchise

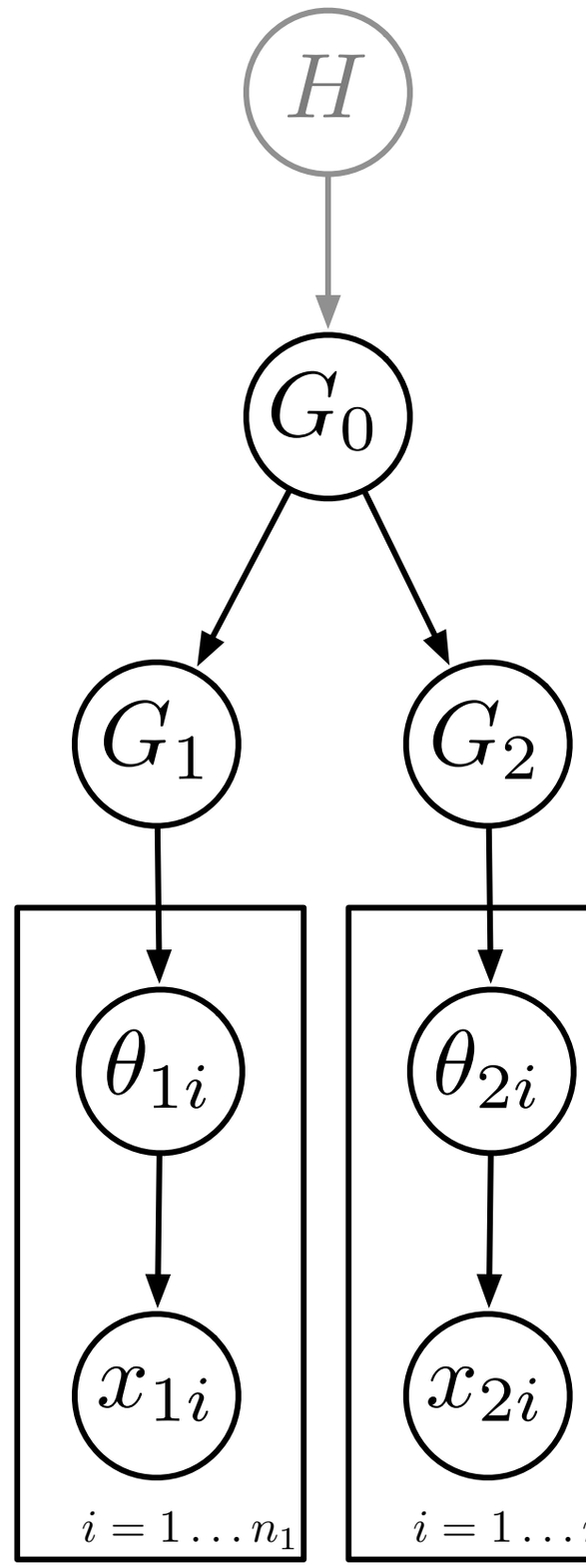




- G_1 and G_2 can both be represented using CRPs.

Chinese Restaurant Franchise

- G_0 can also be represented using a CRP.



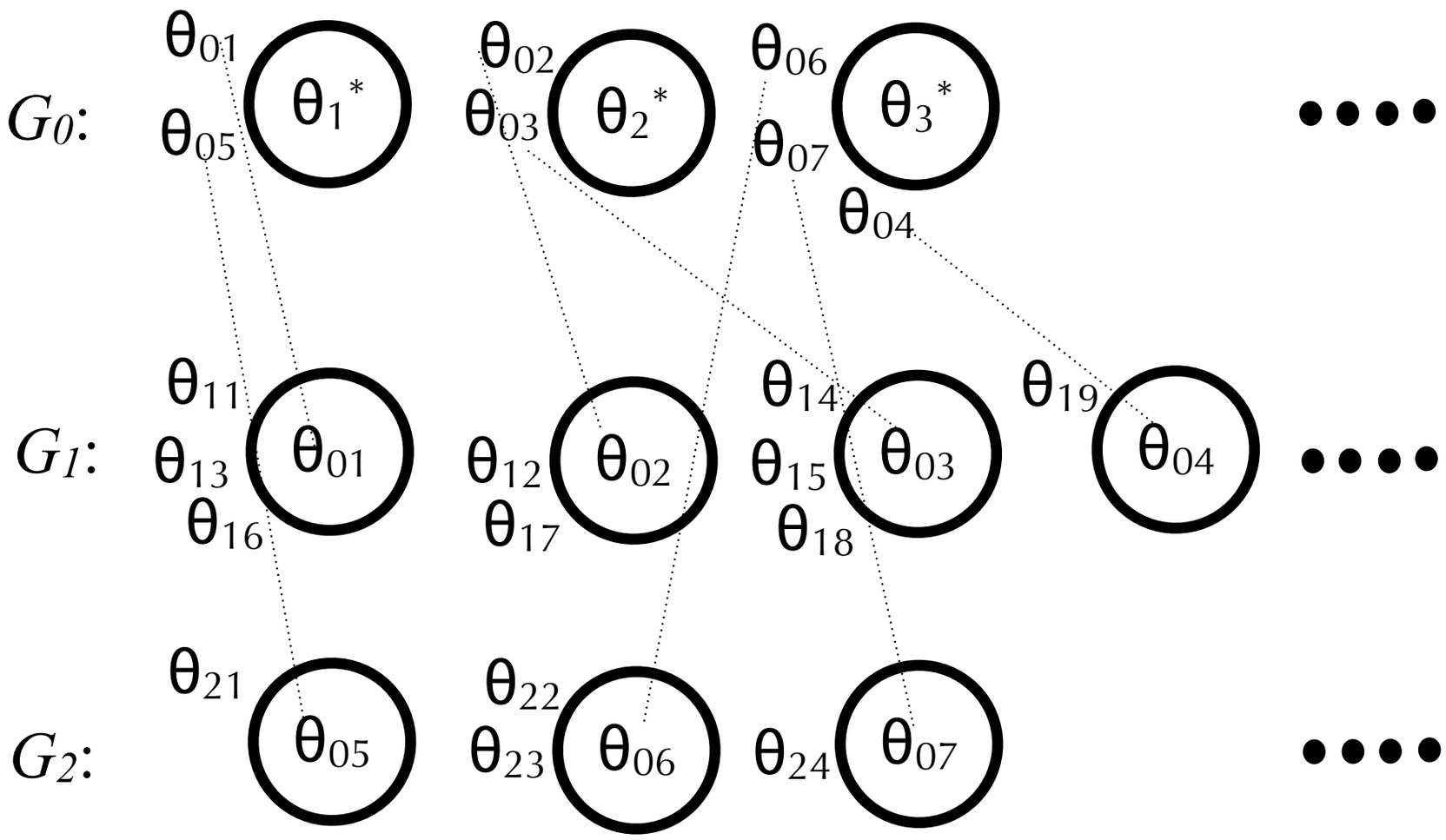
G_1: θ_{11} θ_{13} θ_{16} θ_{01} θ_{12} θ_{17} θ_{02} θ_{14} θ_{15} θ_{18} θ_{03} θ_{19} θ_{04} ••••

G_2: θ_{21} θ_{05} θ_{22} θ_{23} θ_{06} θ_{24} θ_{07} ••••

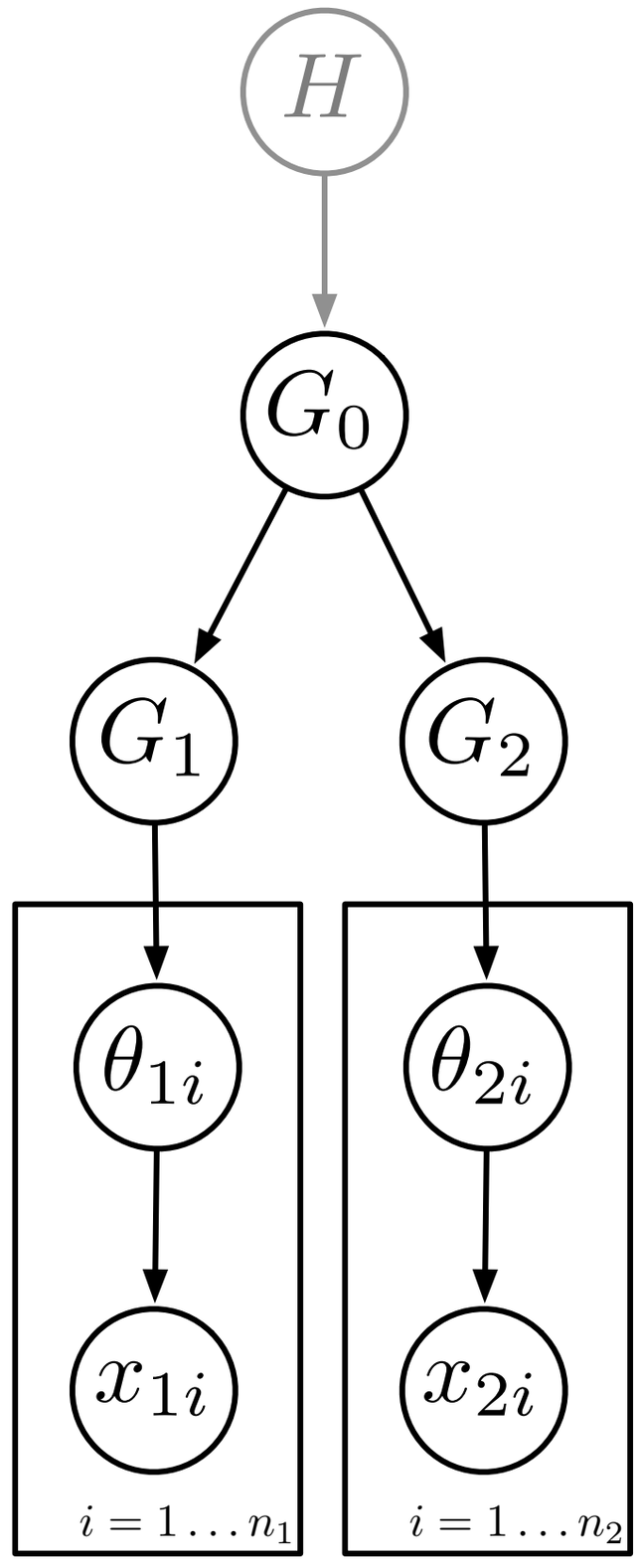
- G_1 and G_2 can both be represented using CRPs.

Chinese Restaurant Franchise

- G_0 can also be represented using a CRP.

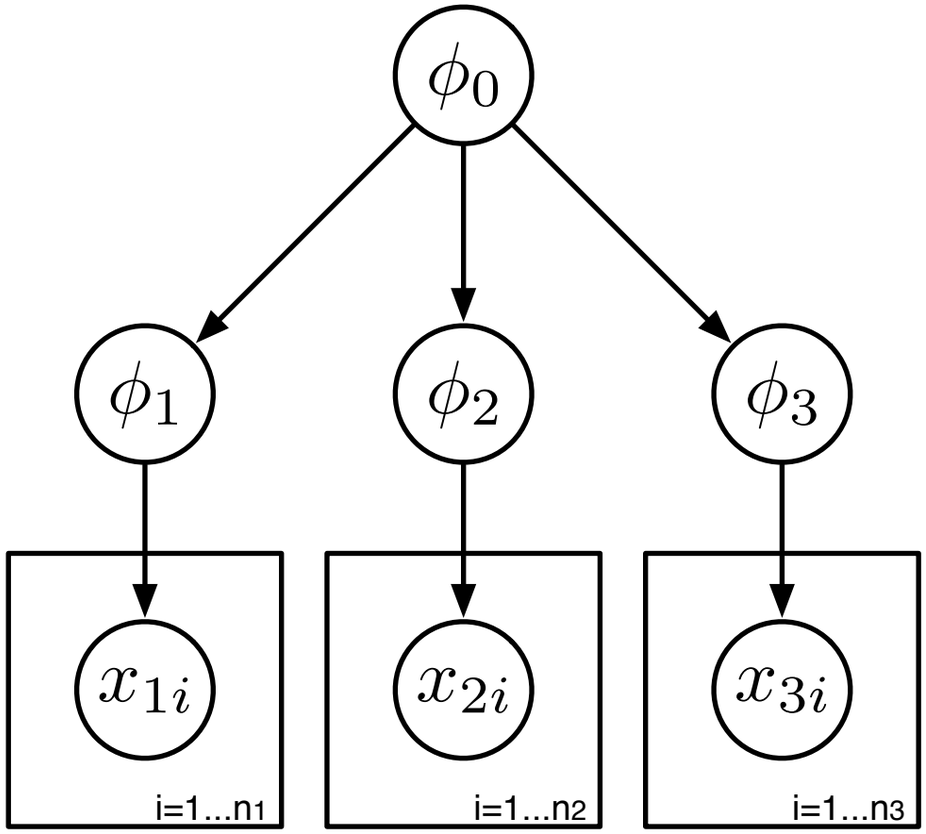


- G_1 and G_2 can both be represented using CRPs.



Hierarchical Bayesian Modelling

- An important overarching theme in modern statistics.
- In machine learning, have been used for multitask learning, transfer learning, learning-to-learn and domain adaptation.



[Gelman et al, 1995, James & Stein 1961]

Hierarchical Bayesian Nonparametrics

- Bayesian nonparametric models are increasingly used as building blocks by modellers to build complex probabilistic models.
- Hierarchical modelling are a natural technique for combining building blocks.
- Applications span computational linguistics, time series and sequential models, vision, genetics etc.

- Dependent random measures:
  - techniques for introducing dependencies among random measures indexed by spatial or temporal covariates.
- Nested processes:
  - technique for modelling heterogeneity in data.

Dependent Random Measures

- A measure-valued stochastic process $\{G_\phi\}$ indexed by a covariate space Φ.
- G_ϕ is the random measure at location $\phi \in \Phi$.
- If each G_ϕ is marginally DP, we have a **dependent Dirichlet process**.

- Density regression: estimating density over output space conditional on ϕ.
- Applications include image segmentation, topic models through time, dictionary learning, spatial models, and many others in biostatistics, signal processing etc.

[MacEachern 1999, Dunson 2010]

Hierarchical Pitman-Yor Language Model

[Goldwater et al 2006, Teh 2006]

n-gram Language Models

Sequence Models for Language and Text

- Probabilistic models for sequences of words and characters, e.g.

south, parks, road

s, o, u, t, h, _, p, a, r, k, s, _, r, o, a, d

- ***n*-gram language models** are high order Markov models of such discrete sequence:

$$P(\text{sentence}) = \prod_i P(\text{word}_i | \text{word}_{i-N+1} \ldots \text{word}_{i-1})$$

n-gram Language Models

- High order Markov models:

$$P(\text{sentence}) = \prod_i P(\text{word}_i | \text{word}_{i-N+1} \dots \text{word}_{i-1})$$

- Large vocabulary size means naïvely estimating parameters of this model from data counts is problematic for N>2.

$$P^{\text{ML}}(\text{word}_i | \text{word}_{i-N+1} \dots \text{word}_{i-1}) = \frac{C(\text{word}_{i-N+1} \dots \text{word}_i)}{C(\text{word}_{i-N+1} \dots \text{word}_{i-1})}$$

- Naïve priors/regularization fail as well: most parameters have *no* associated data.
  - Smoothing.
  - Hierarchical Bayesian models.

Smoothing in Language Models

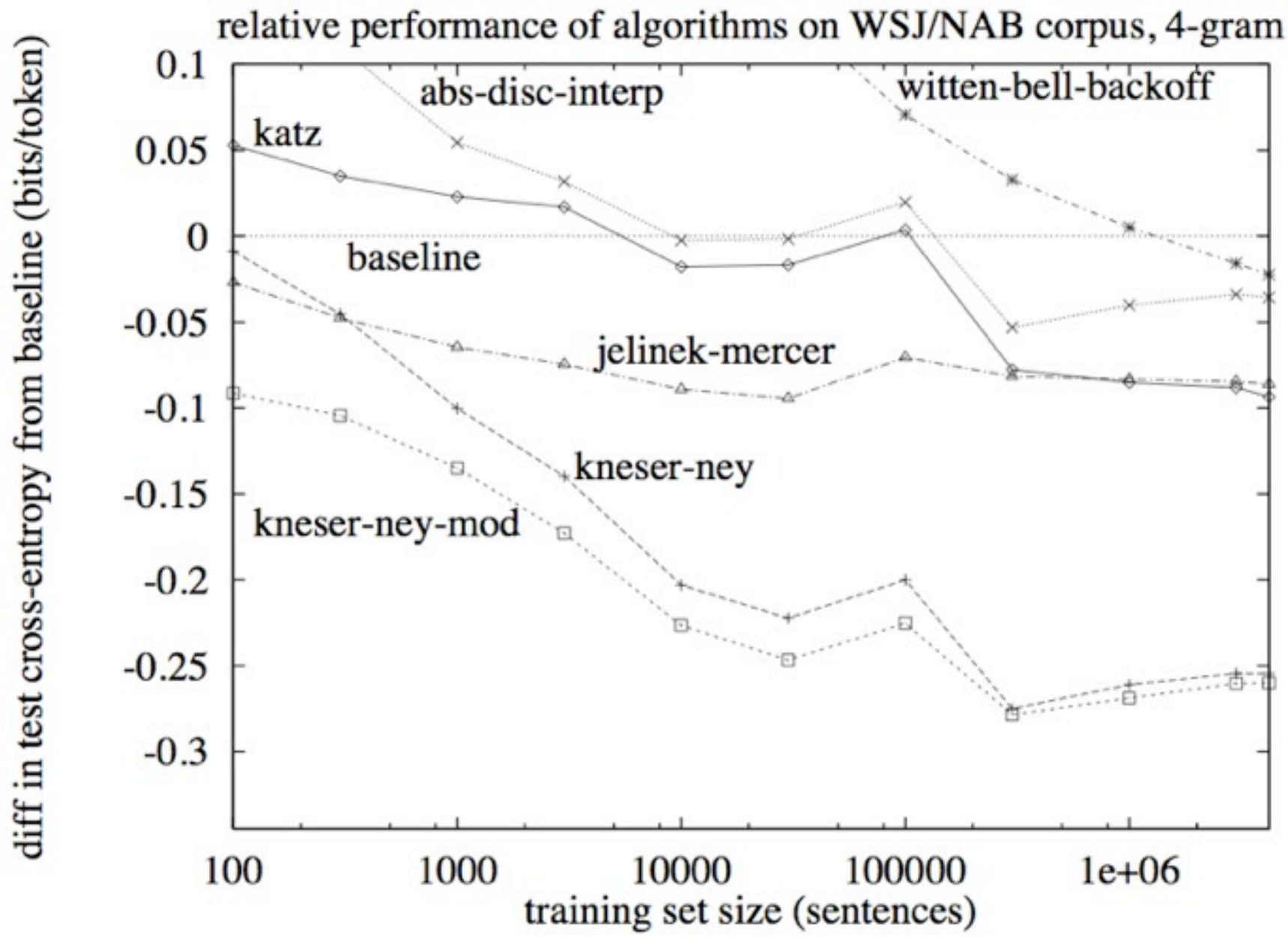
- **Smoothing** is a way of dealing with data sparsity by combining large and small models together.

$$P^{\text{smooth}}(\text{word}_i|\text{word}_{i-N+1}^{i-1}) = \sum_{n=1}^{N} \lambda(n) Q_n(\text{word}_i|\text{word}_{i-n+1}^{i-1})$$

- Combines expressive power of large models with better estimation of small models (cf bias-variance trade-off).

$$\begin{aligned} & P^{\text{smooth}}(\text{road}|\text{south parks}) \\ = \quad & \lambda(3) Q_3(\text{road}|\text{south parks}) + \\ & \lambda(2) Q_2(\text{road}|\text{parks}) + \\ & \lambda(1) Q_1(\text{road}|\emptyset) \end{aligned}$$

Smoothing in Language Models



- Interpolated and modified Kneser-Ney are best.

[Chen & Goodman 1999]

Hierarchical Pitman-Yor Language Models

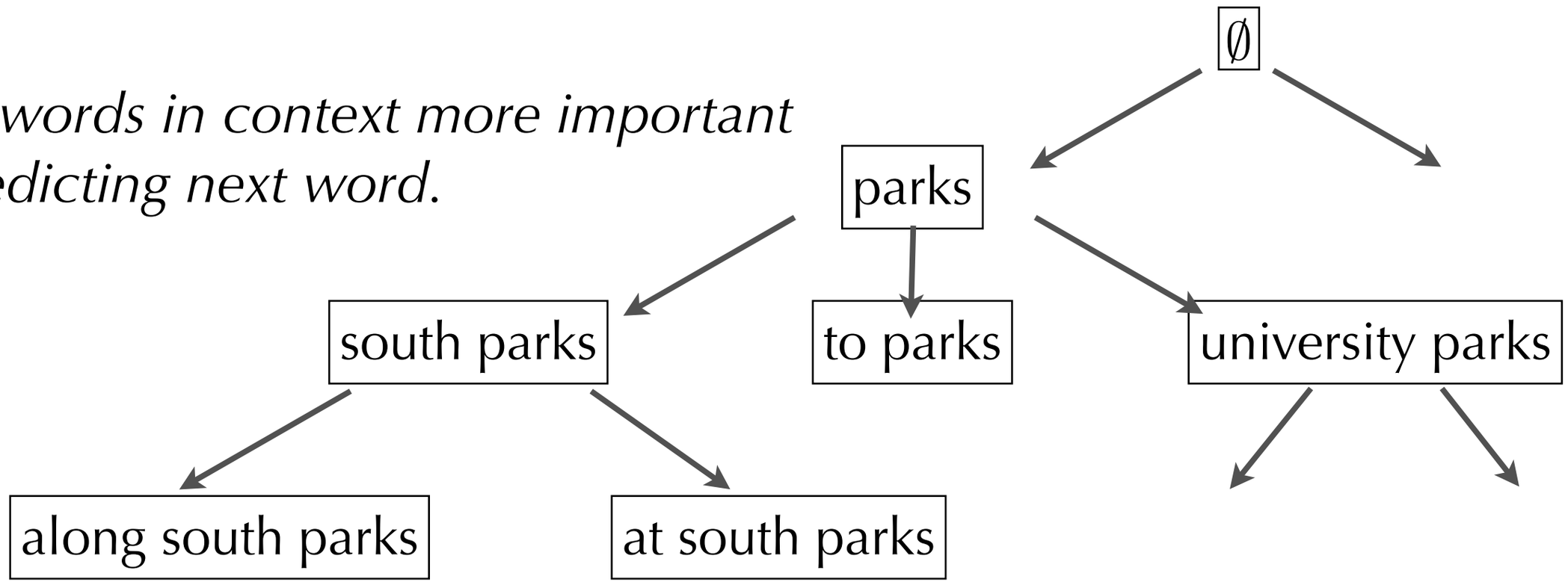
Context Tree

- **Context** of conditional probabilities naturally organized using a tree.

$$
\begin{aligned}
& P^{\text{smooth}}(\text{road}|\text{south parks}) \\
= \quad & \lambda(3)Q_3(\text{road}|\text{south parks}) + \\
& \lambda(2)Q_2(\text{road}|\text{parks}) + \\
& \lambda(1)Q_1(\text{road}|\emptyset)
\end{aligned}
$$

- Smoothing makes conditional probabilities of neighbouring contexts more similar.

- *Later words in context more important in predicting next word.*



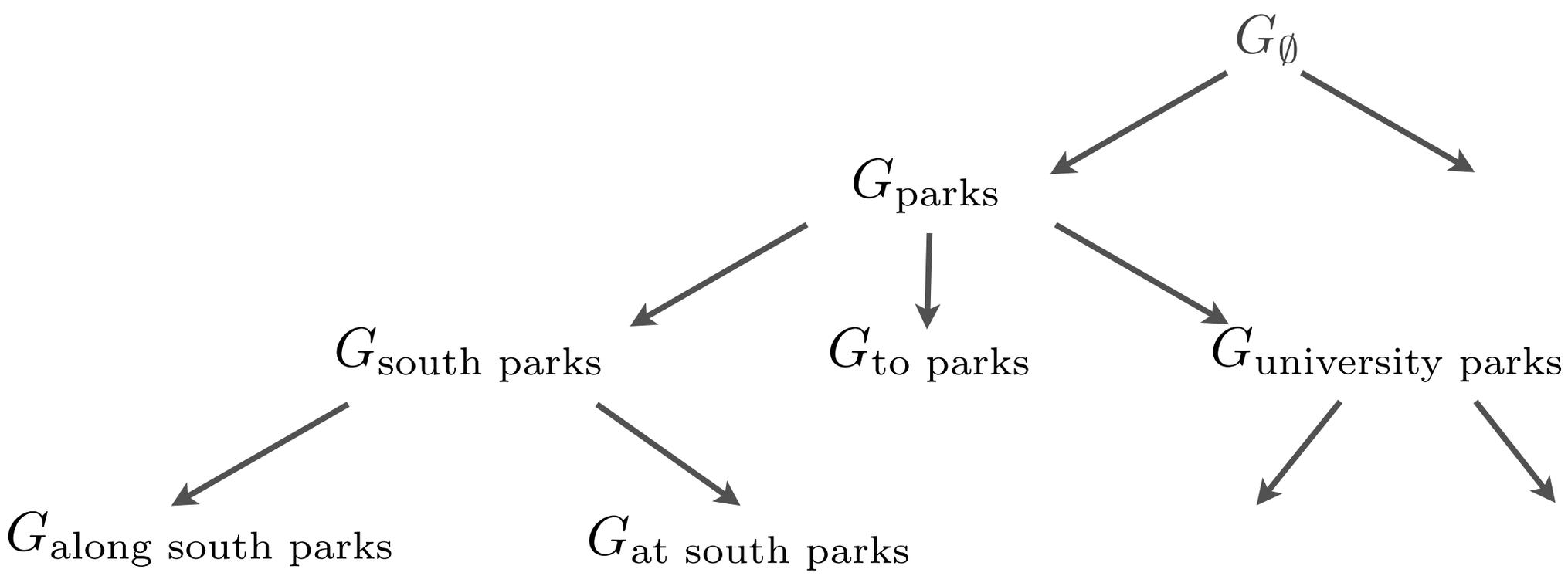
Hierarchical Bayes on Context Tree

- Parametrize the conditional probabilities of Markov model:

$$P(\text{word}_i = w | \text{word}_{i-N+1}^{i-1} = u) = G_u(w)$$

$$G_u = [G_u(w)]_{w \in \text{vocabulary}}$$

- G_u is a probability vector associated with context u.



Hierarchical Dirichlet Language Models

- What is $P(G_u|G_{\mathrm{pa}(u)})$? Obvious choice is the standard Dirichlet distribution over probability vectors.

T	N-1	IKN	MKN	HDLM
2×10^6	2	148.8	144.1	191.2
4×10^6	2	137.1	132.7	172.7
6×10^6	2	130.6	126.7	162.3
8×10^6	2	125.9	122.3	154.7
10×10^6	2	122.0	118.6	148.7
12×10^6	2	119.0	115.8	144.0
14×10^6	2	116.7	113.6	140.5
14×10^6	1	169.9	169.2	180.6
14×10^6	3	106.1	102.4	136.6

- We will use Pitman-Yor processes instead.

[MacKay and Peto 1994]

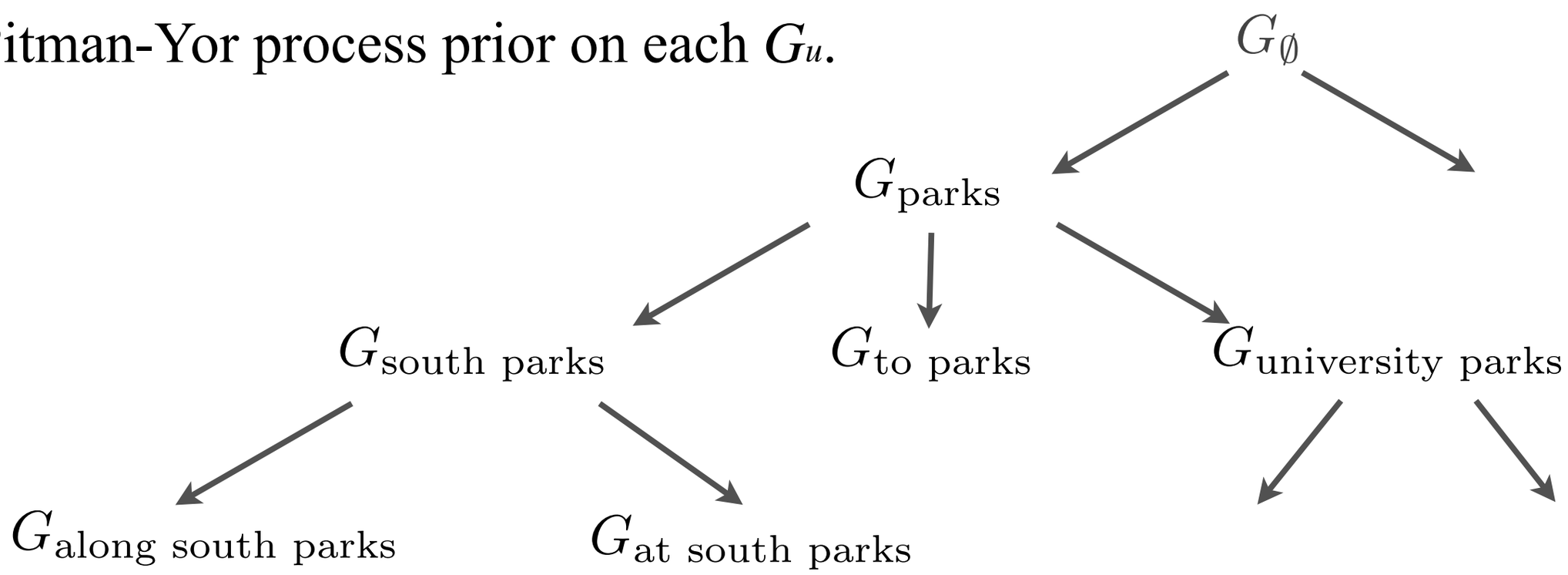
Hierarchical Pitman-Yor Language Models

- Parametrize the conditional probabilities of Markov model:

$$P(\text{word}_i = w | \text{word}_{i-N+1}^{i-1} = u) = G_u(w)$$

$$G_u = [G_u(w)]_{w \in \text{vocabulary}}$$

- G_u is a probability vector associated with context u.
- Place Pitman-Yor process prior on each G_u.



[Goldwater et al 2006, Teh 2006]

Hierarchical Pitman-Yor Language Models

- Significantly improved on the hierarchical Dirichlet language model.
- Results better Kneser-Ney smoothing, state-of-the-art language models.

T	N-1	IKN	MKN	HDLM	HPYLM
2×10^6	2	148.8	**144.1**	191.2	144.3
4×10^6	2	137.1	**132.7**	172.7	**132.7**
6×10^6	2	130.6	126.7	162.3	**126.4**
8×10^6	2	125.9	122.3	154.7	**121.9**
10×10^6	2	122.0	118.6	148.7	**118.2**
12×10^6	2	119.0	115.8	144.0	**115.4**
14×10^6	2	116.7	113.6	140.5	**113.2**
14×10^6	1	169.9	**169.2**	180.6	169.3
14×10^6	3	106.1	102.4	136.6	**101.9**

- Similarity of perplexities not a surprise---Kneser-Ney can be derived as a particular approximate inference method.

Hierarchical Pitman-Yor Process

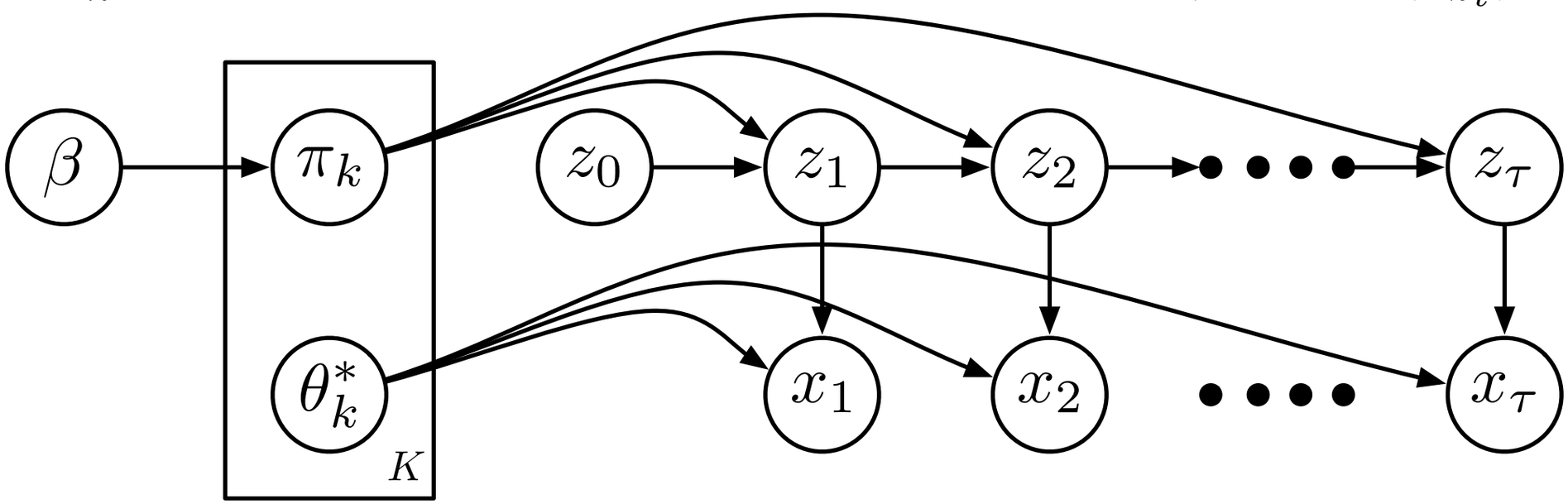
- Application of hierarchical Pitman-Yor processes to *n*-gram language models:
  - Hierarchical Bayesian modelling allows for sharing of statistical strength and improved parameter estimation.
  - Pitman-Yor processes has power law properties more suitable in modelling linguistic data.
- State-of-the-art language models, theoretical justification for another state-of-the-art model called interpolated Kneser-Ney.

Infinite Hidden Markov Model

Hidden Markov Models

$$\pi_k \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \qquad z_t|z_{t-1} \sim \pi_{z_{t-1}}$$

$$\theta_k^* \sim H \qquad x_t|z_t \sim H(\theta_{z_t}^*)$$

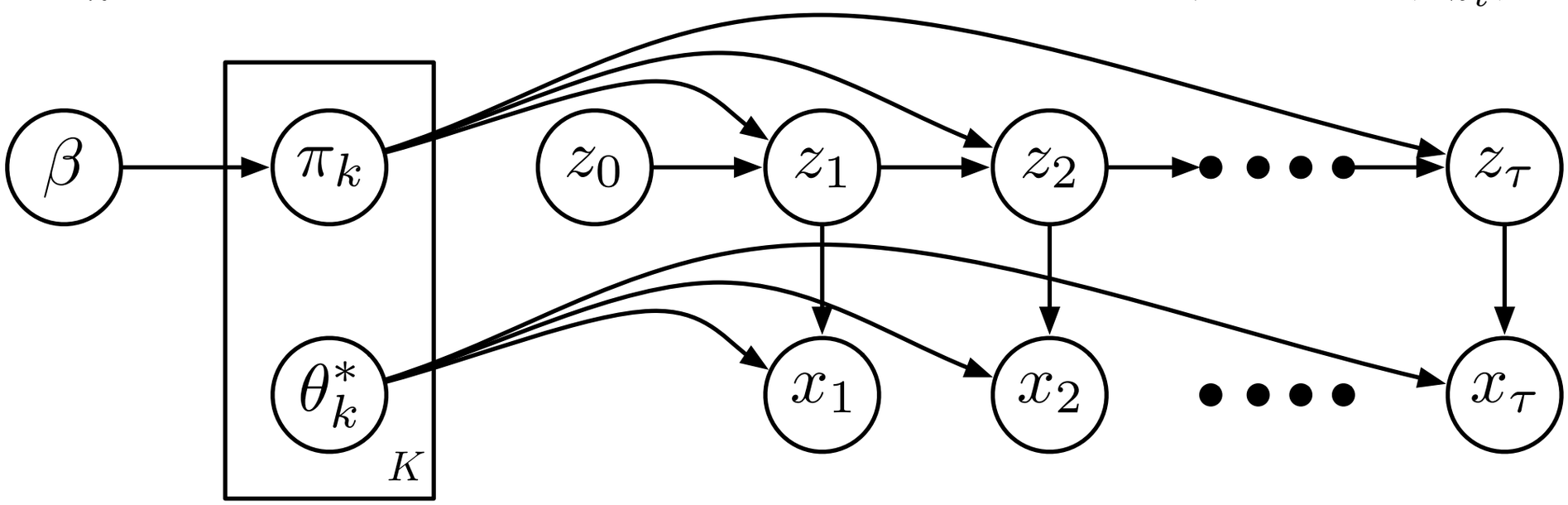


- Can we take $K \to \infty$?

Infinite Hidden Markov Models

$$\pi_k \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \qquad z_t|z_{t-1} \sim \pi_{z_{t-1}}$$

$$\theta_k^* \sim H \qquad x_t|z_t \sim H(\theta_{z_t}^*)$$



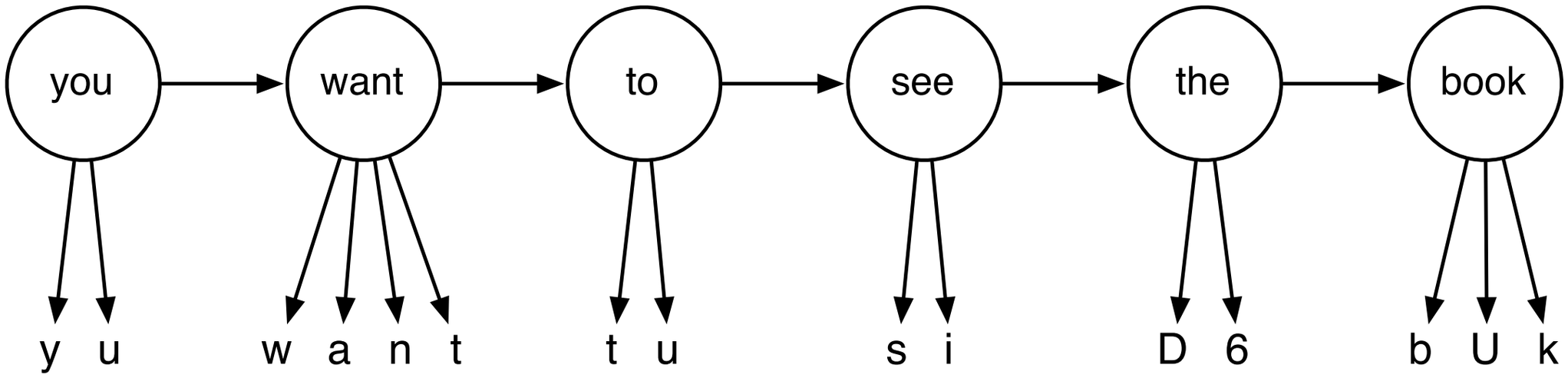
- Cannot simply take $K \to \infty$ for the model above; same failure as LDA.
- Again can use a hierarchical Dirichlet process to define an **infinite hidden Markov model**.

[Beal et al 2002, Teh et al 2006]

Word Segmentation

- 山花 貞夫 ・ 新 民連 会長 は 十六 日 の 記者 会見 で 、 村山 富市 首相 ら 社会党 執行 部 と さきがけ が 連携 強化 を めざした 問題 に ついて 「 私 たち の 行動 が 新しい 政界 の 動き を 作った と いえる 。 統一 会派 を 超え て 将来 の 日本 の ...
- 今后 一段 时期 , 不但 居民 会 更 多 地 选择 国债 , 而且 一些 金融 机构 在 准备金 利率 调 低 后 , 出于 安全性 方面 的 考虑 , 也 会 将 部分 资金 用来 购买 国债 。
- yuwanttusiD6bUk?

iHMM Word Segmentation



yuwanttusiD6bUk

- Number of word types is unknown (and part of the output of learning).
- We can use the infinite HMM coupled with a model to generate strings of characters for each word.

[Goldwater et al 2006, Mochihashi et al 2009]

iHMM Word Segmentation

	P	R	F	BP	BR	BF	LP	LR	LF
NGS-u	67.7	70.2	68.9	80.6	84.8	82.6	52.9	51.3	52.0
MBDP-1	67.0	69.4	68.2	80.3	84.3	82.3	53.6	51.3	52.4
DP	61.9	47.6	53.8	92.4	62.2	74.3	57.0	57.5	57.2
NGS-b	68.1	68.6	68.3	81.7	82.5	82.1	54.5	57.0	55.7
HDP	**79.4**	**74.0**	**76.6**	**92.4**	**83.5**	**87.7**	**67.9**	**58.9**	**63.1**

Model	MSR	CITYU	Kyoto
NPY(2)	80.2 (51.9)	**82.4** (**126.5**)	62.1 (23.1)
NPY(3)	**80.7** (**48.8**)	81.7 (128.3)	**66.6** (**20.6**)
ZK08	66.7 (—)	69.2 (—)	—

Coagulations, Fragmentations, and Trees

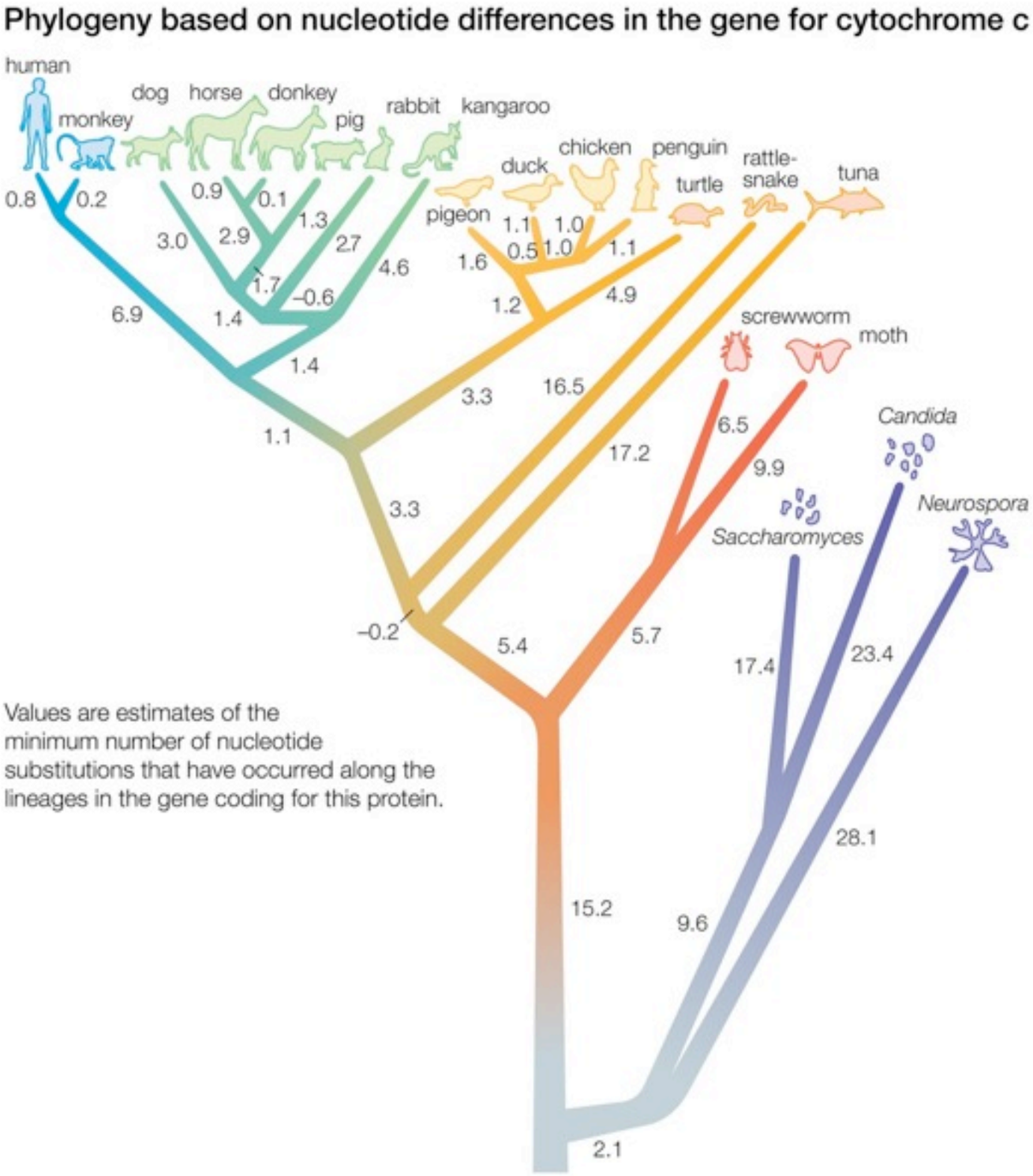
Overview

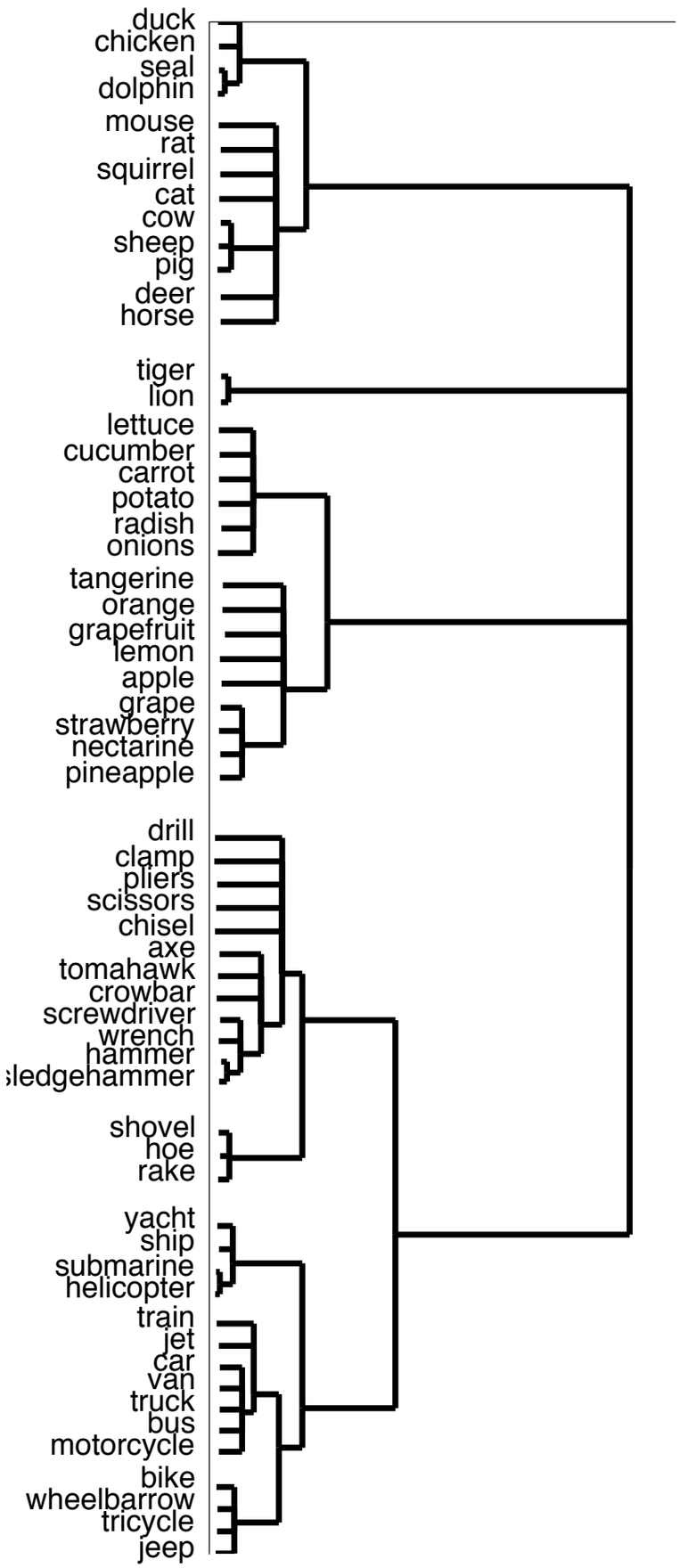
- Bayesian nonparametric learning of trees and hierarchical partitions.
- Fragmentations and coagulations.

- Unifying view of various Bayesian nonparametric models for random trees.

From Random Partitions to Random Trees

Trees


Phylogeny based on nucleotide differences in the gene for cytochrome c
human
monkey
dog
horse
donkey
pig
rabbit
kangaroo
pigeon
duck
chicken
penguin
turtle
rattle-snake
tuna
screwworm
moth
Candida
Neurospora
Saccharomyces
0.8
0.2
0.9
0.1
3.0
2.9
1.3
2.7
1.7
–0.6
4.6
6.9
1.4
1.4
1.1
1.6
1.1
0.5
1.0
1.0
1.1
1.2
4.9
3.3
16.5
17.2
6.5
9.9
3.3
–0.2
5.4
5.7
17.4
23.4
28.1
15.2
9.6
2.1
Values are estimates of the minimum number of nucleotide substitutions that have occurred along the lineages in the gene coding for this protein.


duck
chicken
seal
dolphin
mouse
rat
squirrel
cat
cow
sheep
pig
deer
horse
tiger
lion
lettuce
cucumber
carrot
potato
radish
onions
tangerine
orange
grapefruit
lemon
apple
grape
strawberry
nectarine
pineapple
drill
clamp
pliers
scissors
chisel
axe
tomahawk
crowbar
screwdriver
wrench
hammer
sledgehammer
shovel
hoe
rake
yacht
ship
submarine
helicopter
train
jet
car
van
truck
bus
motorcycle
bike
wheelbarrow
tricycle
jeep

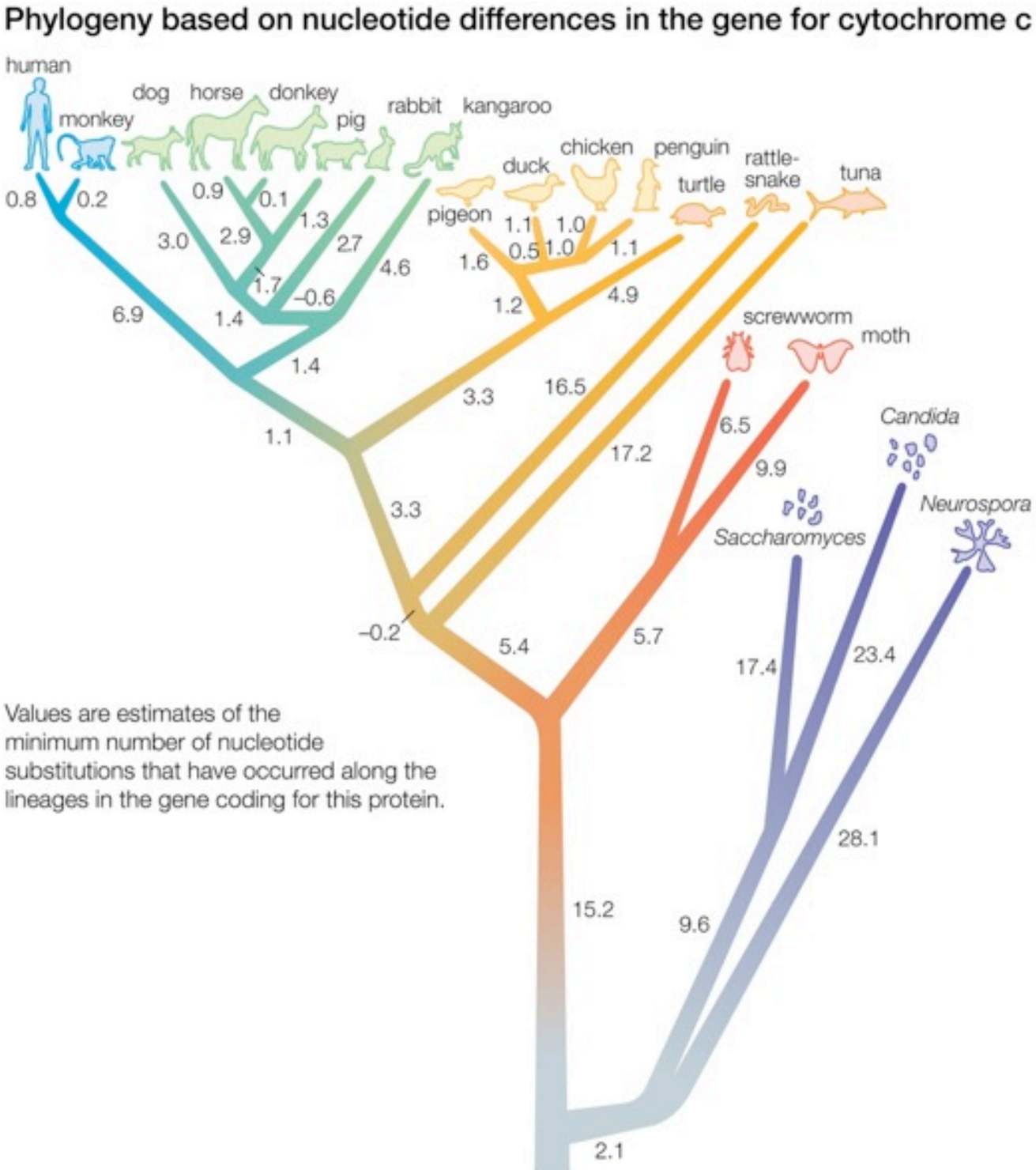
Bayesian Inference for Trees

- Computational and statistical methods for constructing trees:
  - Algorithmic, not model-based.
  - Maximum likelihood
  - Maximum parsimony
- Bayesian inference: introduce prior over trees and compute posterior.

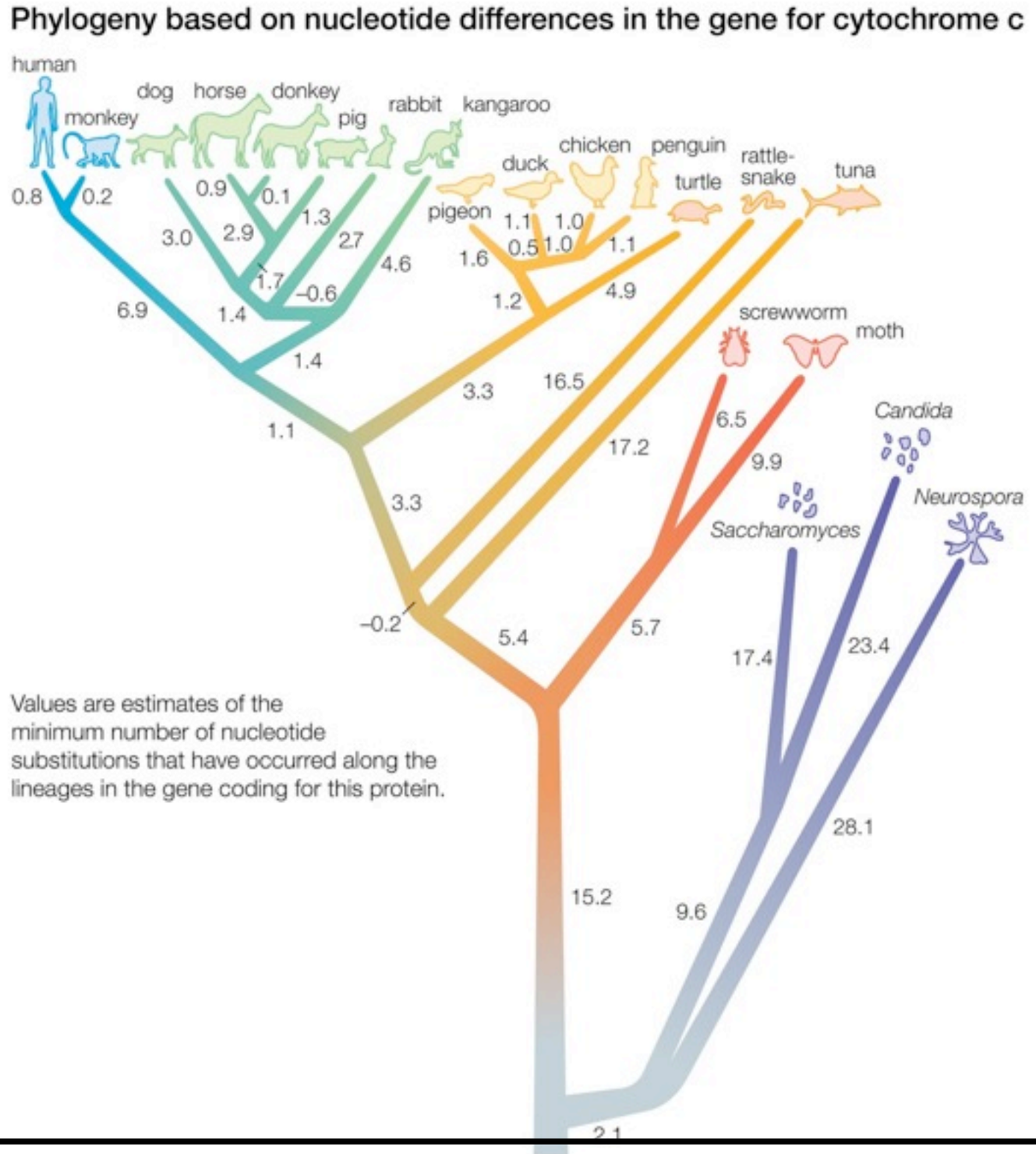
$$P(T|\mathbf{x}) \propto P(T)P(\mathbf{x}|T)$$

- Bayesian nonparametric priors for $P(T)$.
  - Exchangeable and projective models.
- Models for trees has to be nonparametric.

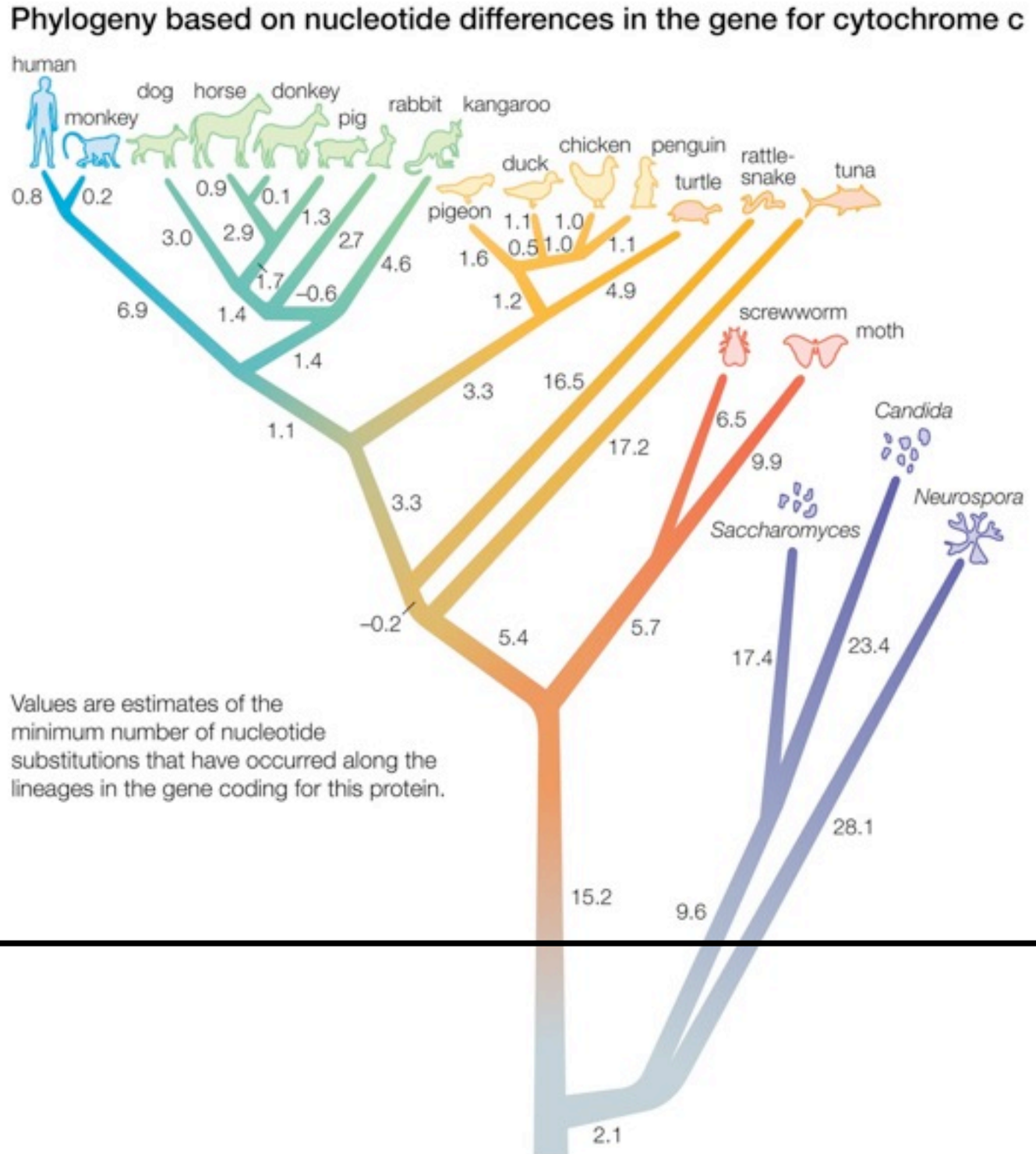
Trees as Sequences of Partitions



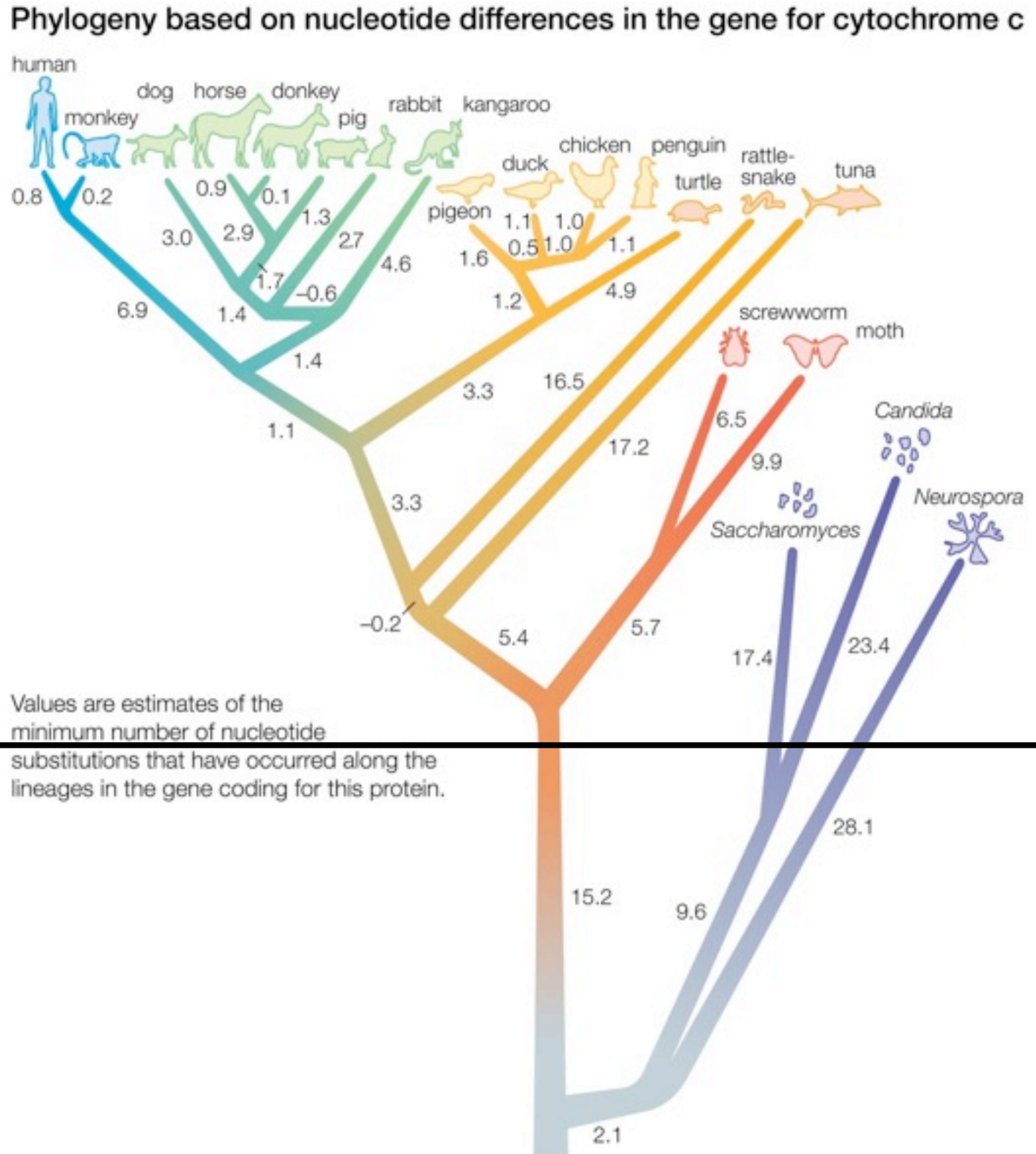
Trees as Sequences of Partitions



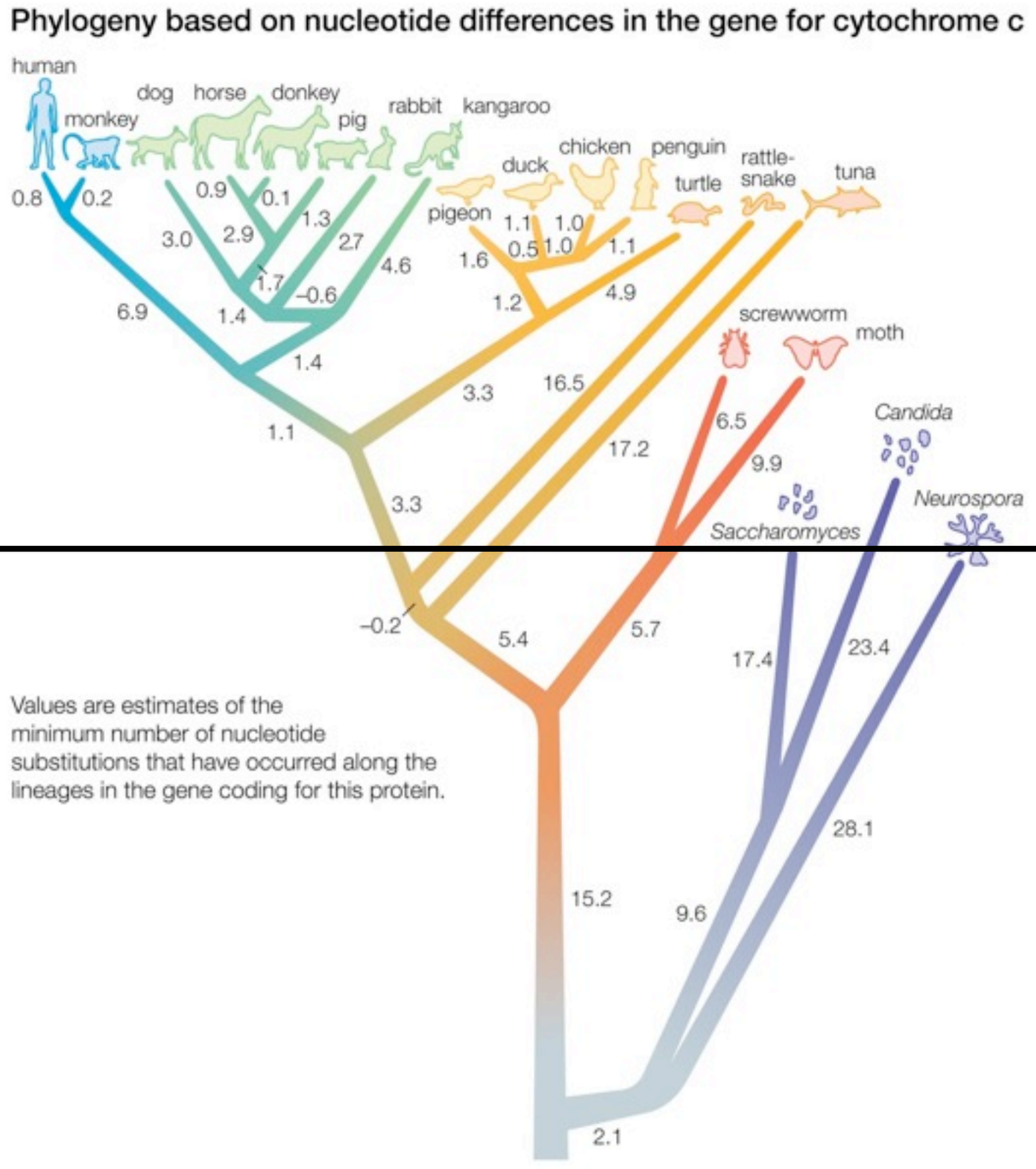
Trees as Sequences of Partitions



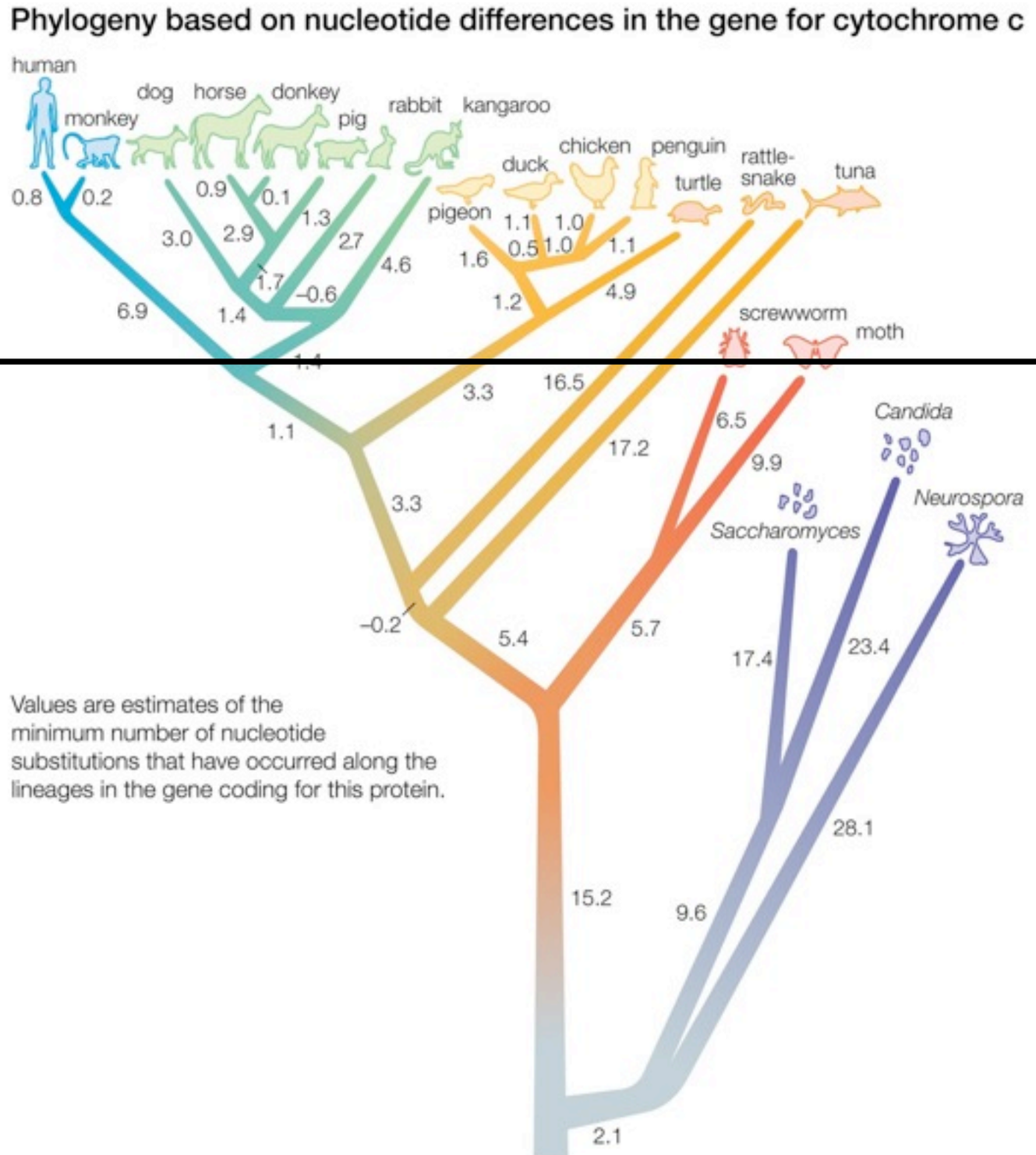
Trees as Sequences of Partitions



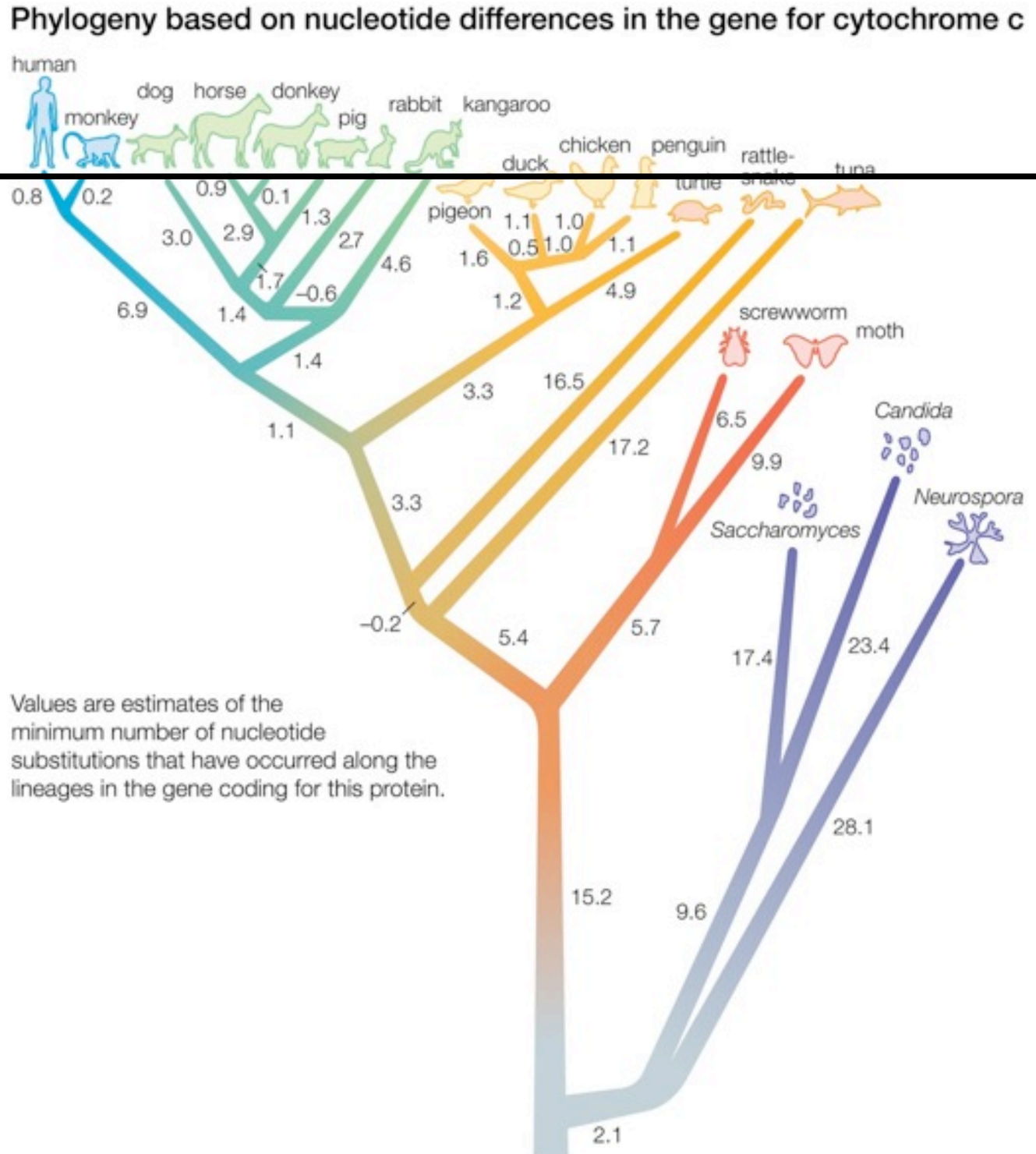
Trees as Sequences of Partitions



Trees as Sequences of Partitions



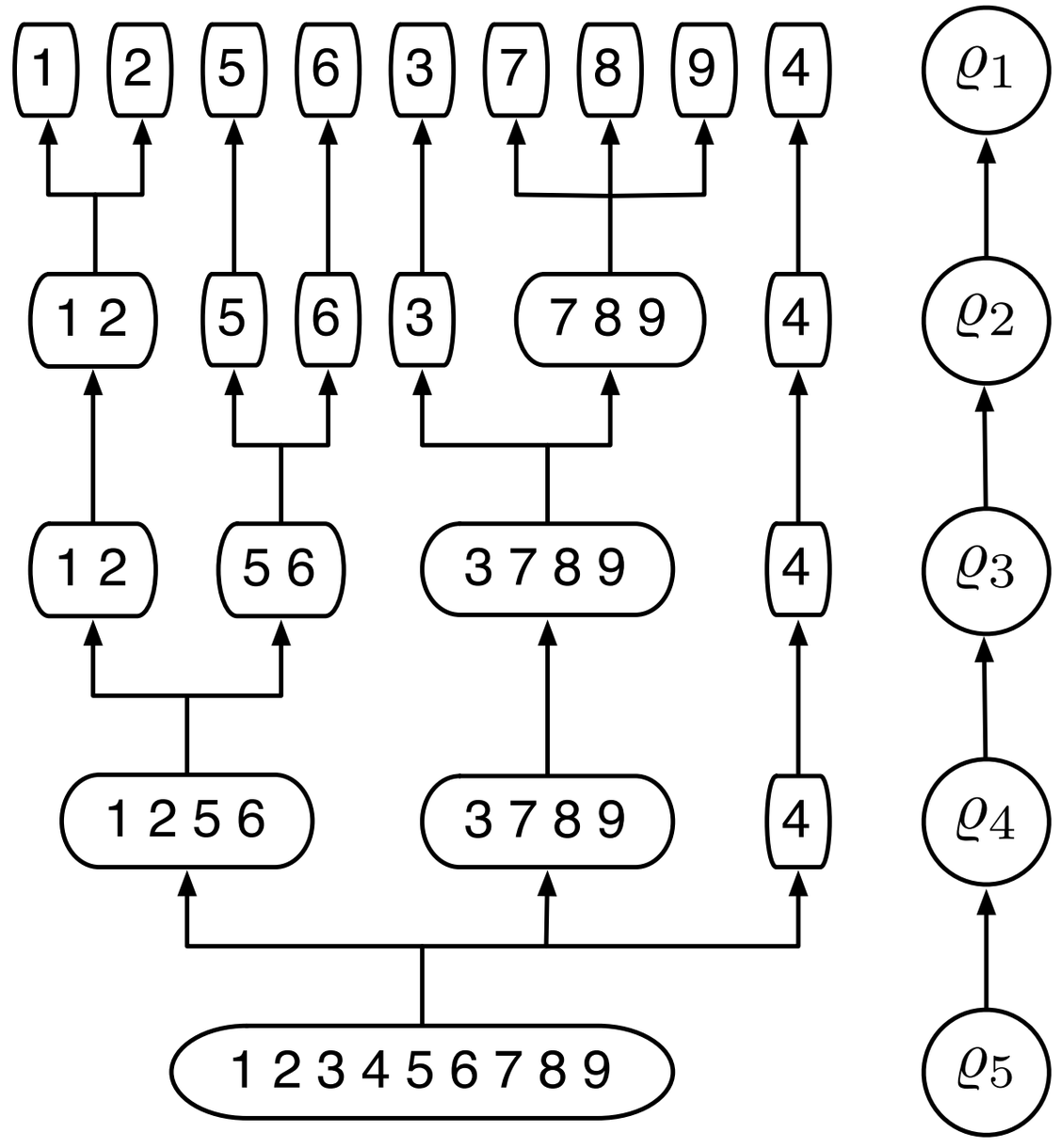
Trees as Sequences of Partitions



Fragmenting Partitions

- Sequence of finer and finer partitions.
- Each cluster fragments until all clusters contain only 1 data item.

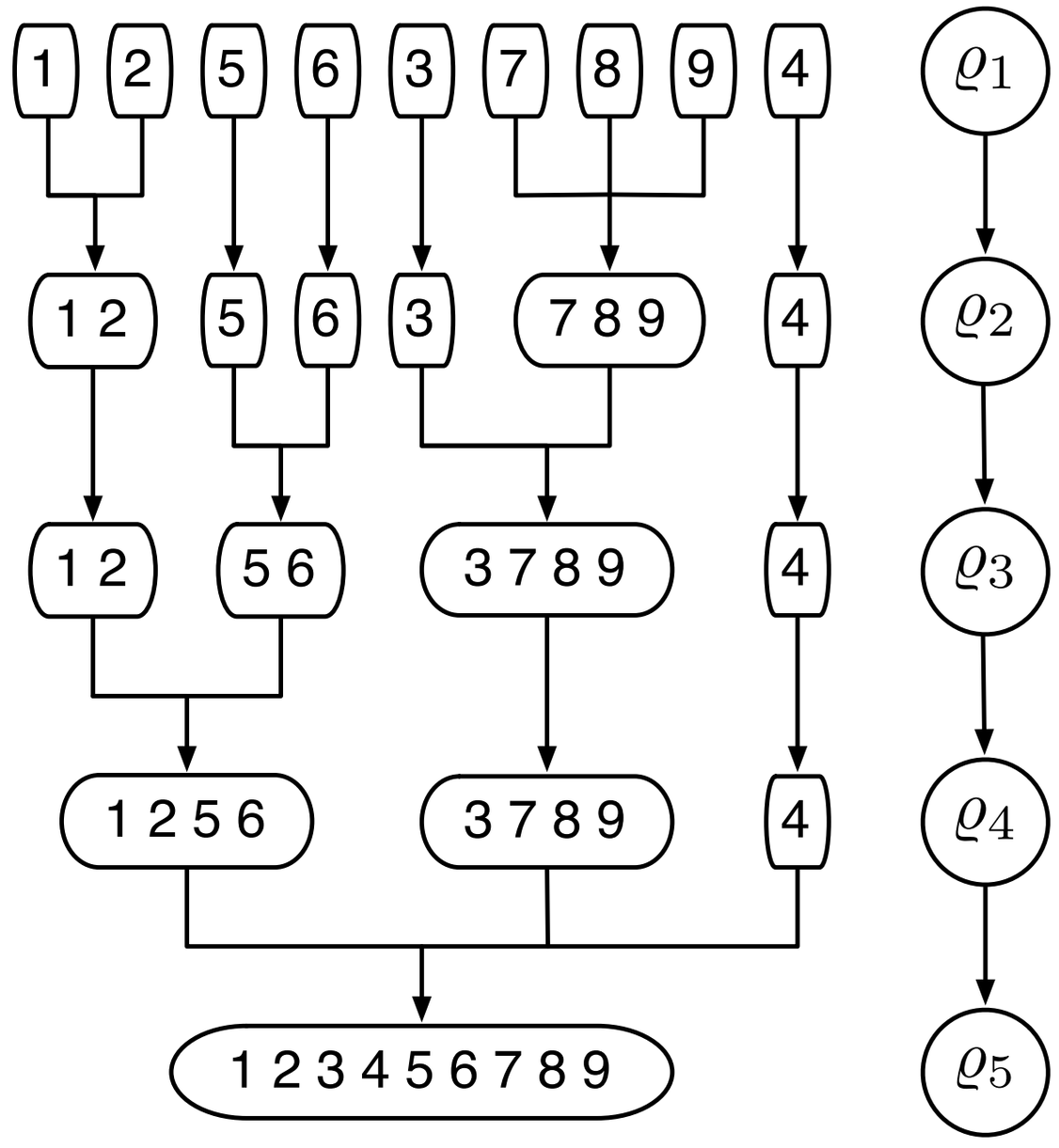
- *Can define a distribution over trees using a Markov chain of fragmenting partitions, with absorbing state* $\mathbf{0}_S$ *(partition where all data items are in their own clusters).*



Coagulating Partitions

- Sequence of coarser and coarser partitions.
- Each cluster formed by coagulating smaller clusters until only 1 left.

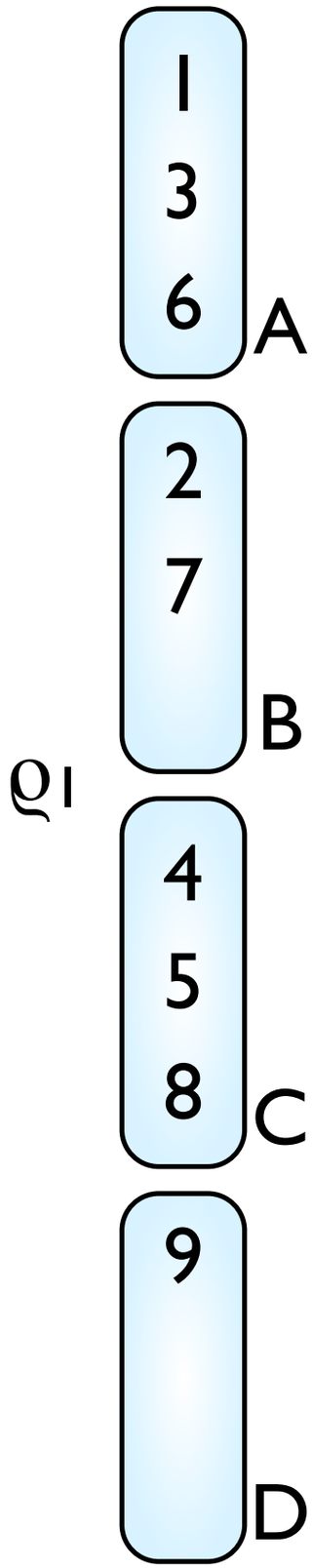
- *Can define a distribution over trees by using a Markov chain of coagulating partitions, with absorbing state* $\mathbf{1}_S$ *(partition where all data items are in one cluster).*



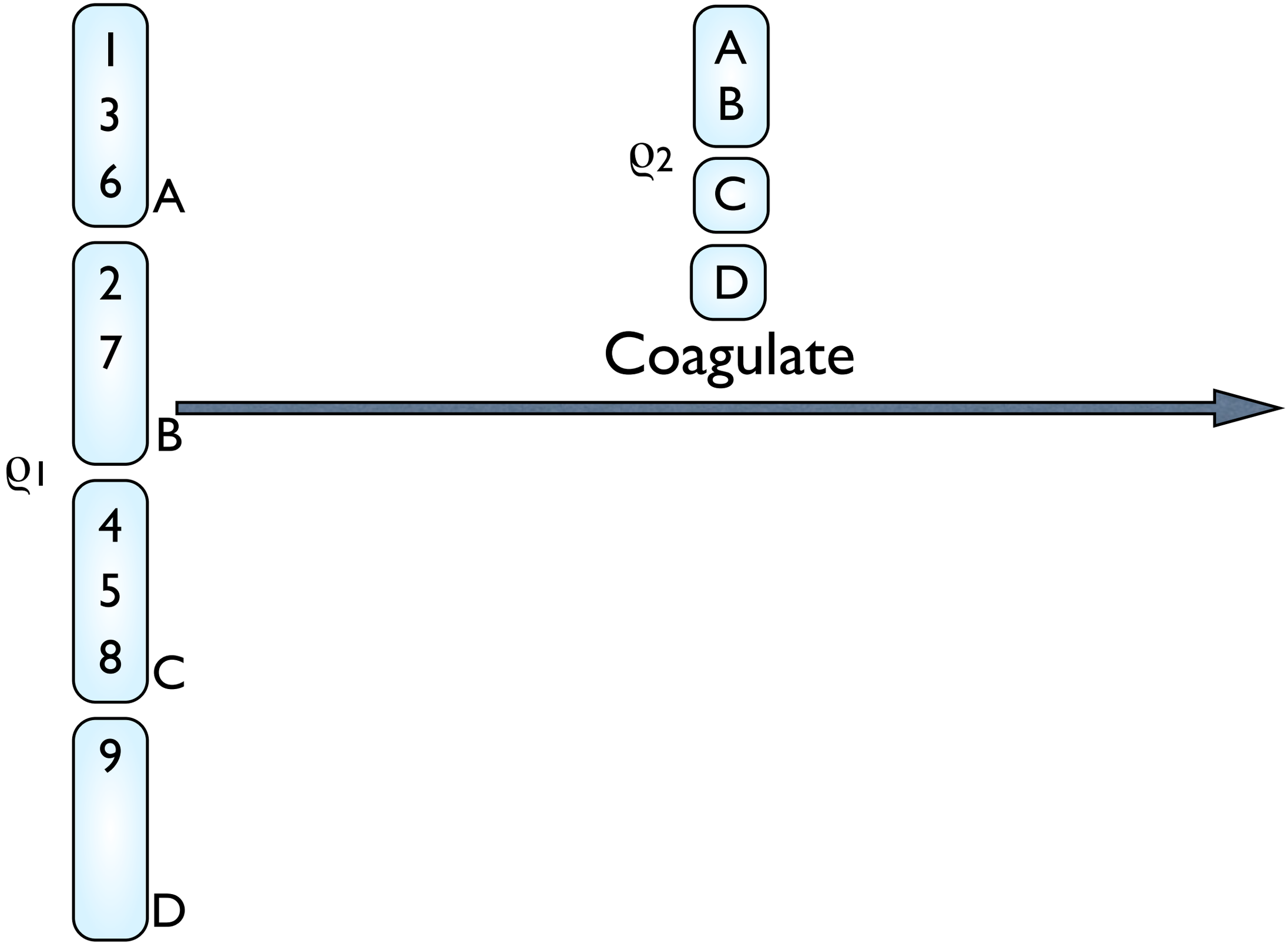
Random Fragmentations and Random Coagulations

[Bertoin 2006]

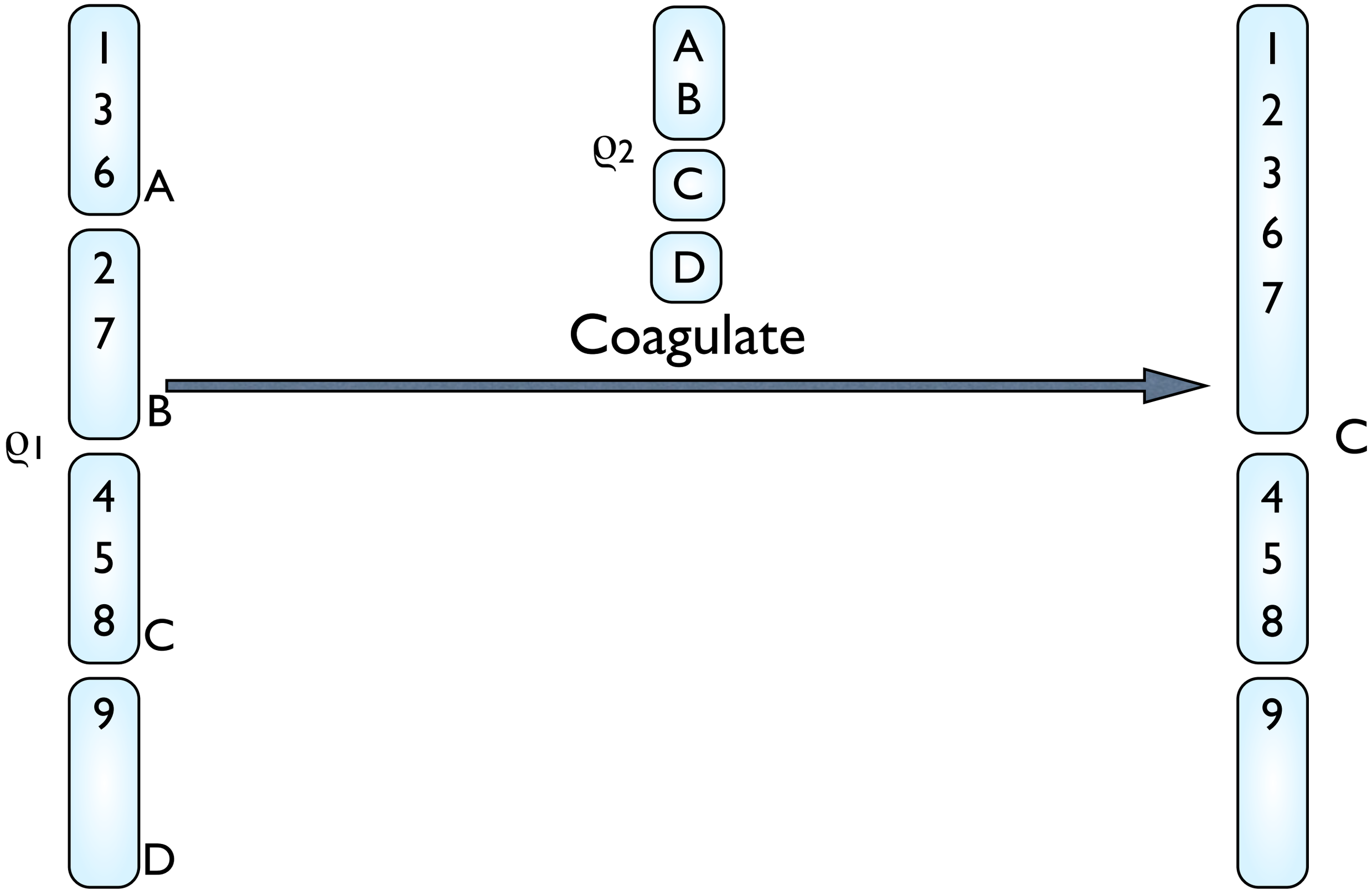
Coagulation and Fragmentation Operators



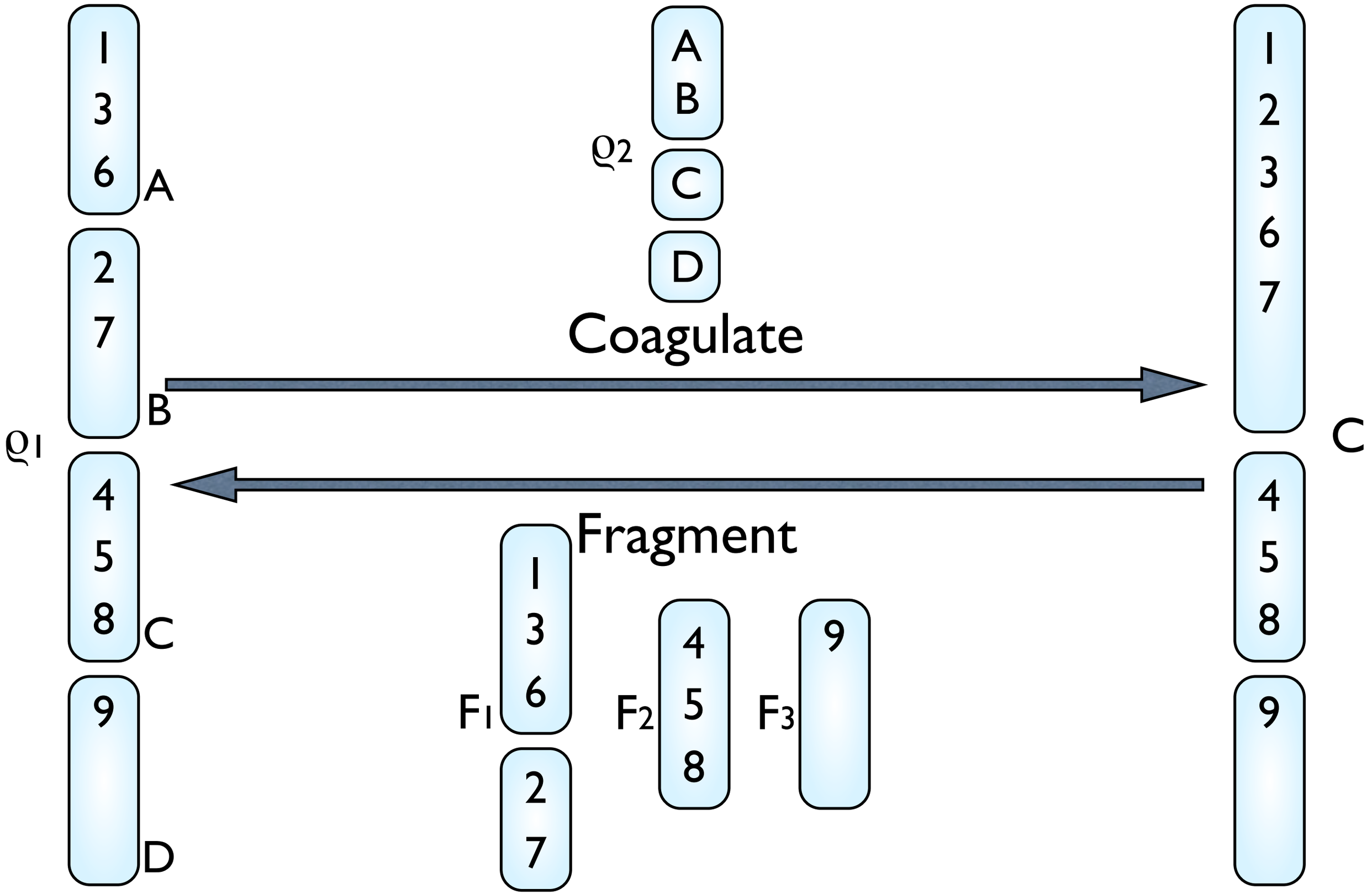
Coagulation and Fragmentation Operators



Coagulation and Fragmentation Operators



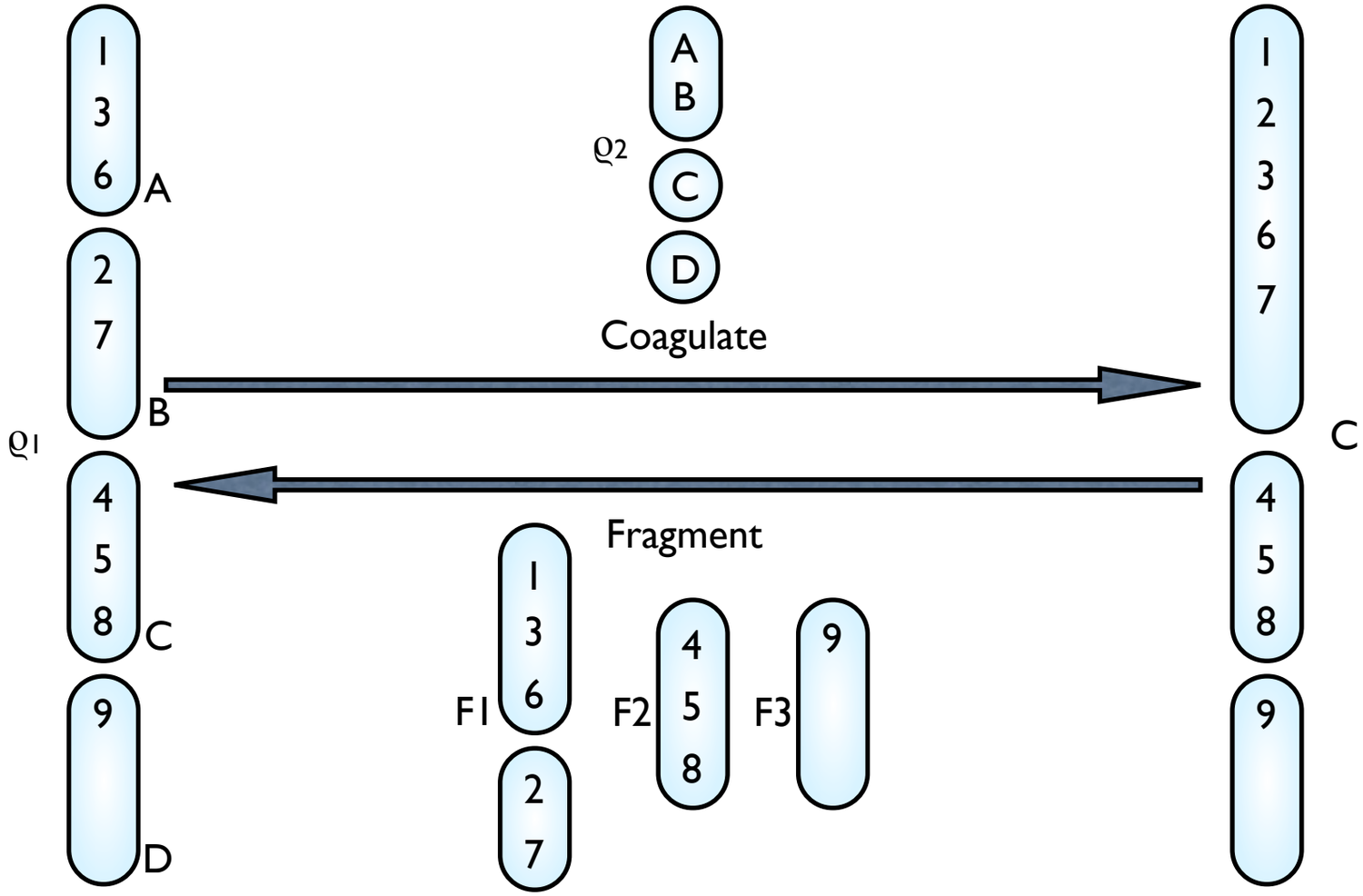
Coagulation and Fragmentation Operators



Random Fragmentations

- Let $C \in \mathcal{P}_{[n]}$ and for each $c \in C$ let $F_c \in \mathcal{P}_c$.
  - Denote **fragmentation** of C by $\{F_c\}$ as $\mathrm{frag}(C,\{F_c\})$.
  - Write $\varrho_1 \mid C \sim \mathrm{FRAG}(C,d,\alpha)$ if $\varrho_1 = \mathrm{frag}(C,\{F_c\})$ with

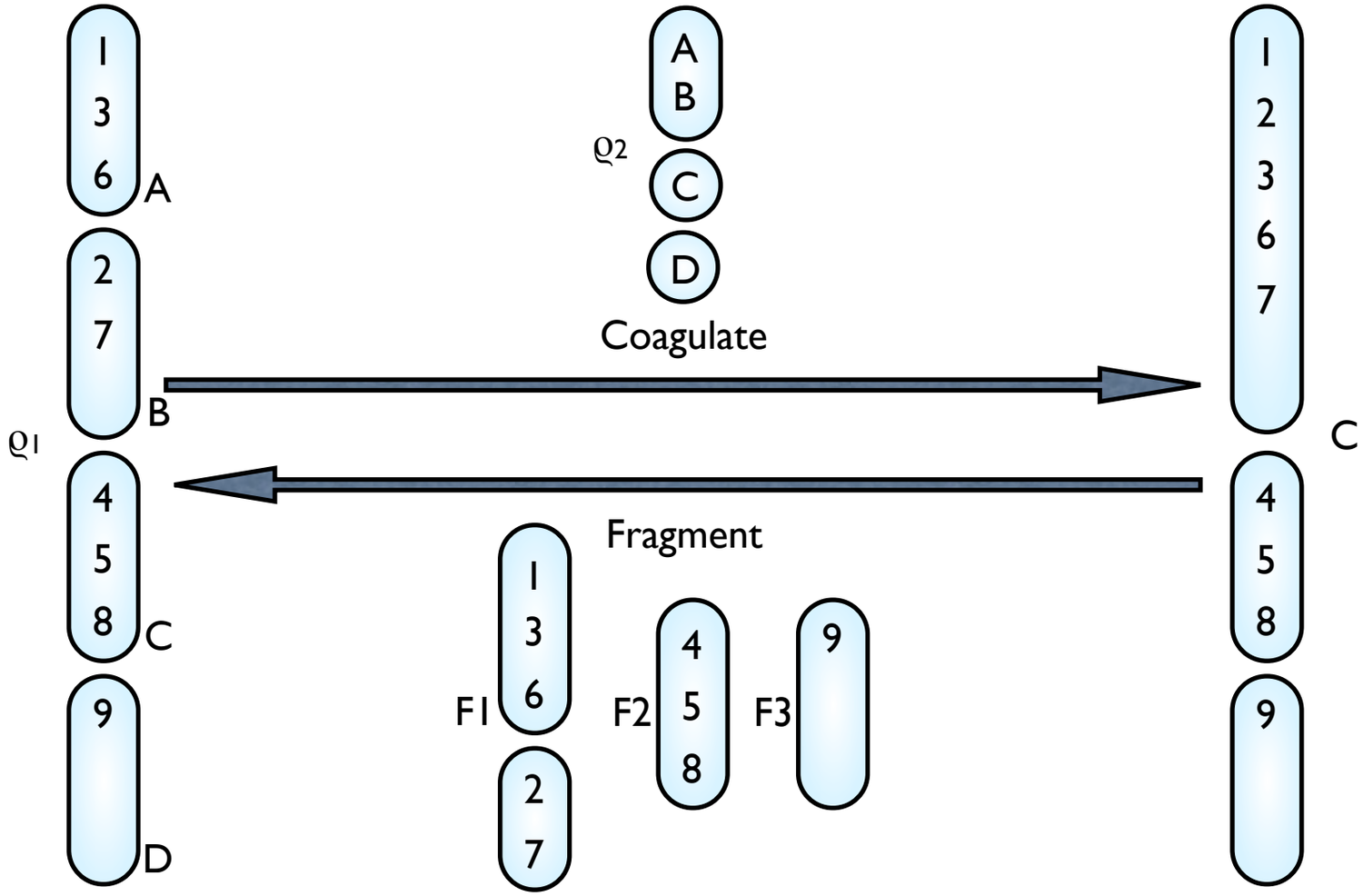
$$F_c \sim \mathrm{CRP}(c,d,\alpha) \text{ independently.}$$



Random Coagulations

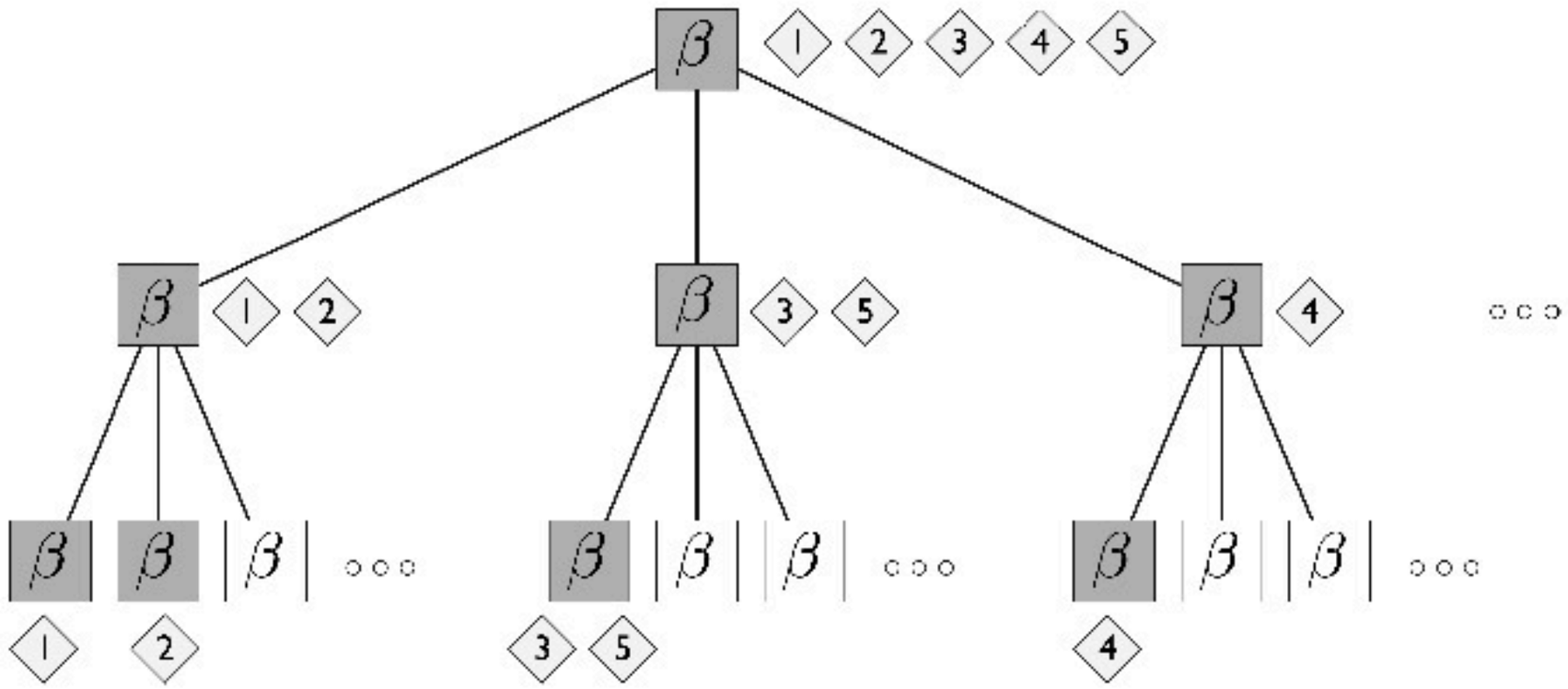
- Let $\varrho_1 \in \mathcal{P}_{[n]}$ and $\varrho_2 \in \mathcal{P}_{\varrho_1}$.
  - Denote **coagulation** of ϱ_1 by ϱ_2 as $\mathrm{coag}(\varrho_1, \varrho_2)$.
  - Write $C \mid \varrho_1 \sim \mathrm{COAG}(\varrho_1, d, \alpha)$ if $C = \mathrm{coag}(\varrho_1, \varrho_2)$ with

$$\varrho_2 \mid \varrho_1 \sim \mathrm{CRP}(\varrho_1, d, \alpha).$$



Random Trees and Random Hierarchical Partitions

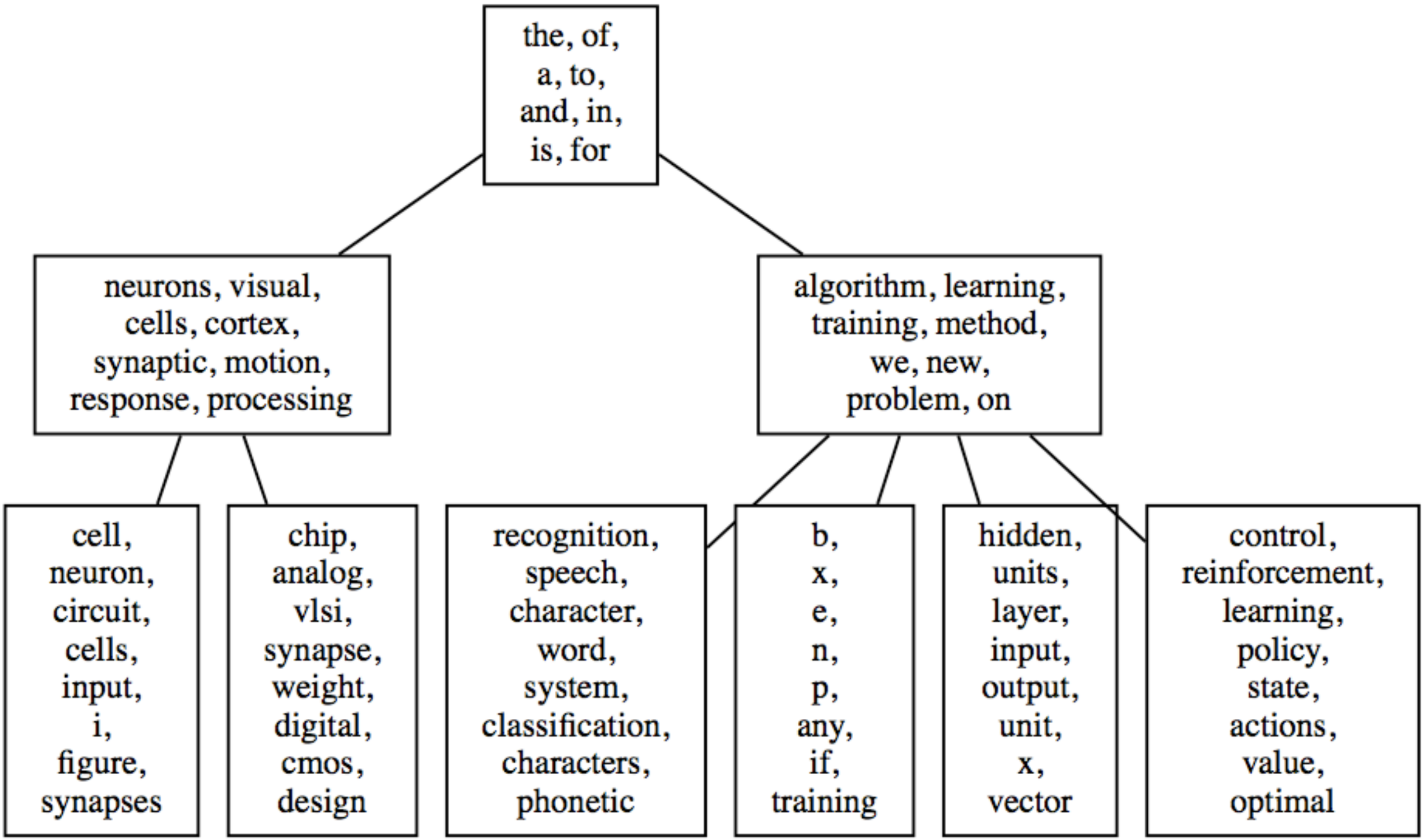
Nested Chinese Restaurant Processes



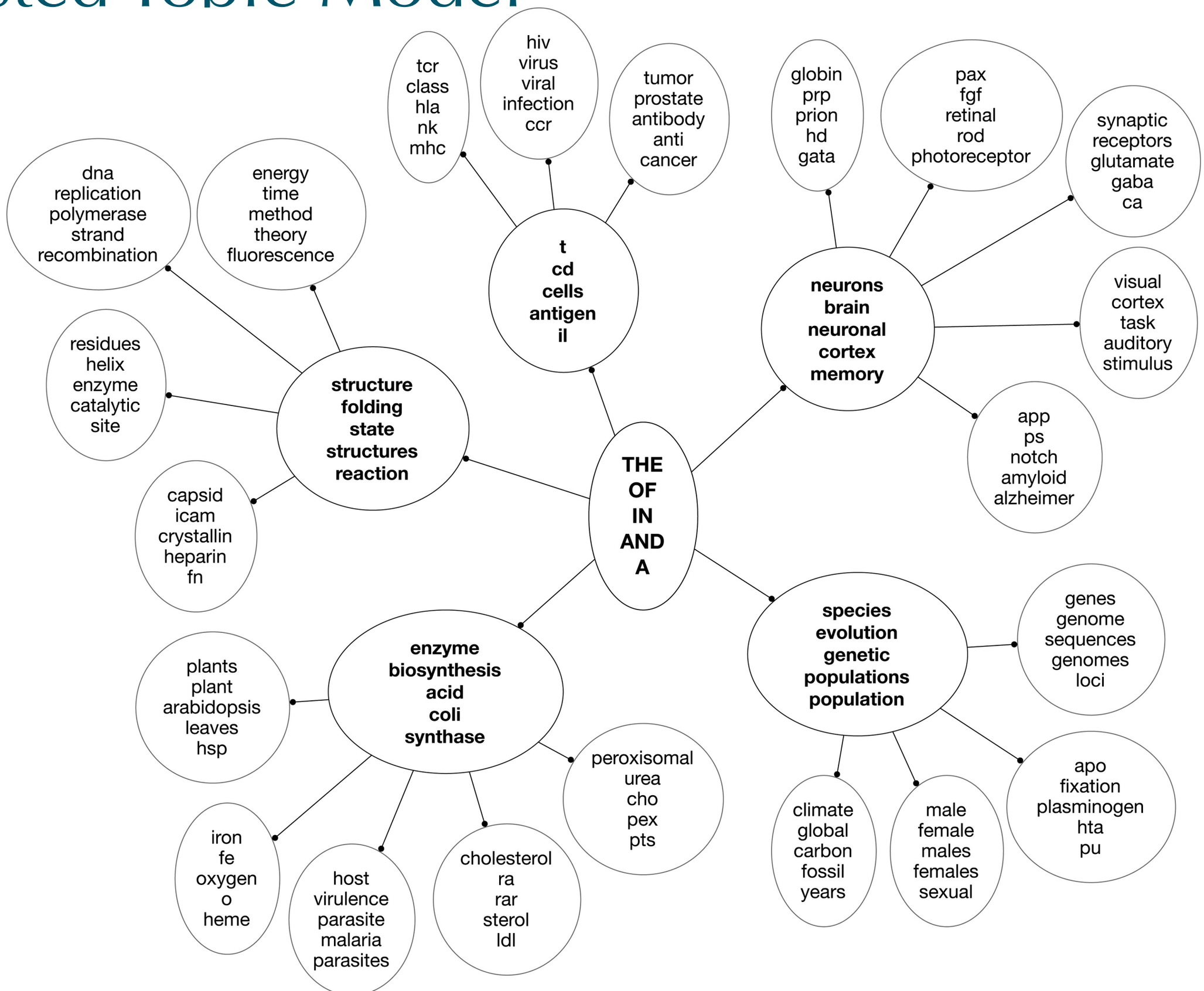
A tourist arrives at the city for an culinary vacation. On the first evening, he enters the root Chinese restaurant and selects a table using the CRP distribution in Eq. (1). On the second evening, he goes to the restaurant identified on the first night's table and chooses a second table using a CRP distribution based on the occupancy pattern of the tables in the second night's restaurant. He repeats this process forever. After M tourists have been on vacation in the city, the collection of paths describes a random subtree of the infinite tree; this subtree has a branching factor of at most M at all nodes. See Figure 3 for an example of the first three levels from such a random tree.

[Blei et al 2004, 2010]

Nested Topic Model



Nested Topic Model

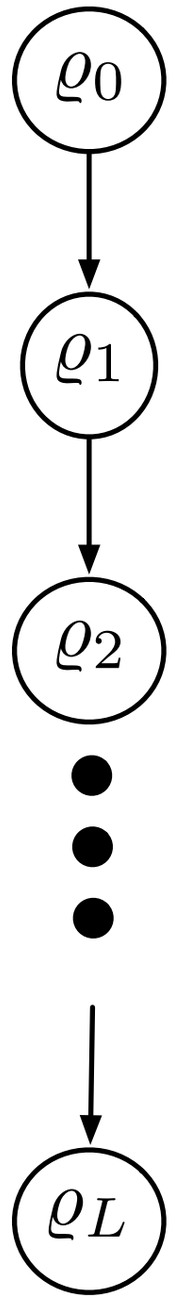


Nested Chinese Restaurant Process

- Start with the null partition $\varrho_0 = \{[n]\}$.
- For each level $l = 1,2,\ldots,L$:

$$\varrho_l = \text{FRAG}(\varrho_{l-1}, 0, \alpha_l)$$

- Fragmentations in different clusters (branches of the hierarchical partition) operate independently.

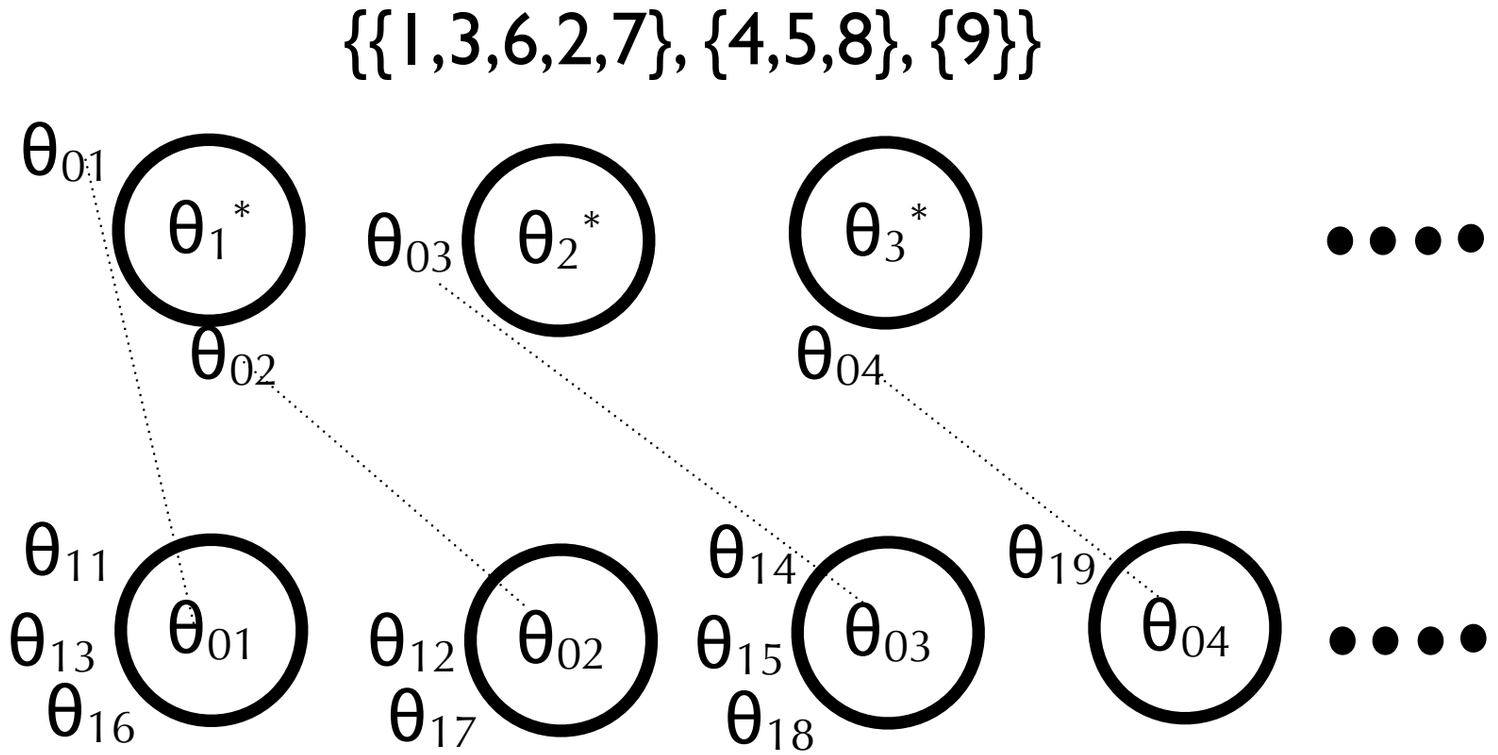


- **Nested Chinese restaurant processes** (nCRP) define a *Markov chain* of partitions, each of which is exchangeable.
- Can be used to define an infinitely exchangeable sequence, with de Finetti measure being the **nested Dirichlet process** (nDP).

[Blei et al 2004, 2010, Rodriguez et al JASA 2008]

Coagulation of Random Partitions

- Consider a Chinese restaurant franchise corresponding to a two level HDP:



$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_1|G_0 \sim \mathrm{DP}(\alpha_1, G_0)$$

- Corresponds to a random coagulation with:

$$\rho_1 \sim \mathrm{CRP}([9], 0, \alpha_1)$$
$$\rho_0|\rho_1 \sim \mathrm{COAG}(\rho_1, 0, \alpha_0)$$

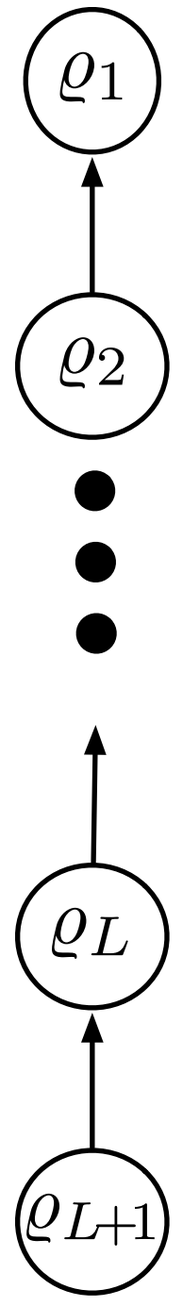
[Teh et al 2006]

Chinese Restaurant Franchise

- For a simple linear hierarchy of DPs (restaurants linearly chained together), the **Chinese restaurant franchise** (CRF) is a sequence of coagulations:
  - At the lowest level $L+1$, we start with the trivial partition $\varrho_{L+1} = \{\{1\},\{2\},...,\{n\}\}$.
  - For each level $l = L,L-1,...,1$:

$$\varrho_l = \mathrm{COAG}(\varrho_{l+1}, 0, \alpha_l)$$

- This is also Markov chain of partitions.

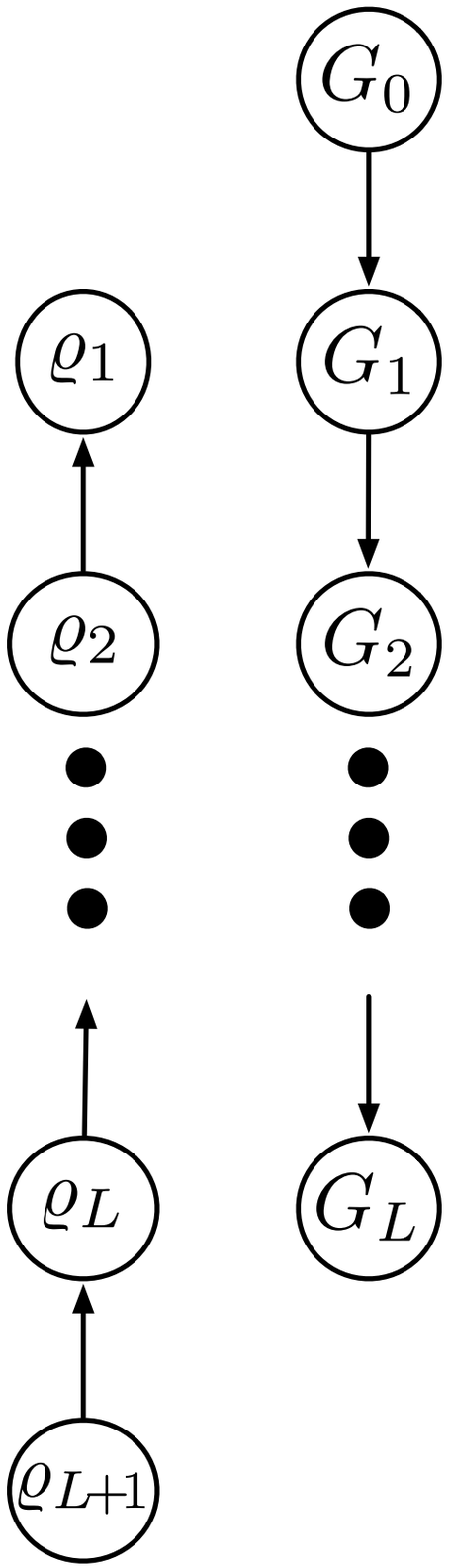


Hierarchical Dirichlet/Pitman-Yor Processes

- Each partition in the Chinese restaurant franchise is again exchangeable.
- The corresponding de Finetti measure is a **Hierarchical Dirichlet process** (HDP).

$$G_l | G_{l-1} \sim \mathrm{DP}(\alpha_l, G_{l-1})$$

- The CRF has been rarely used as a model of hierarchical partitions. Typically it is only used as a convenient representation for inference in the HDP and HPYP.



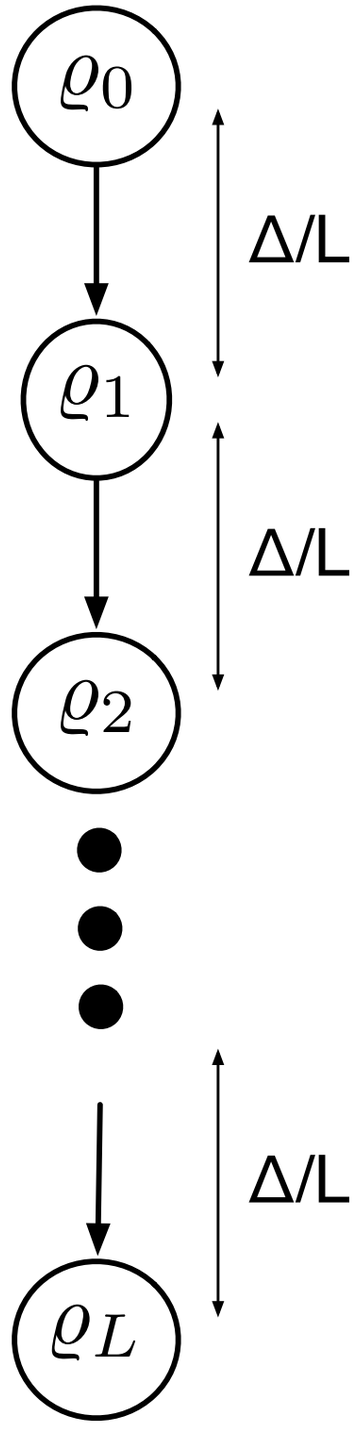
Random Trees

- Nonparametric models of trees are natural.
- Construction of random trees as Markov chains of random partitions.
- Models are infinitely exchangeable.

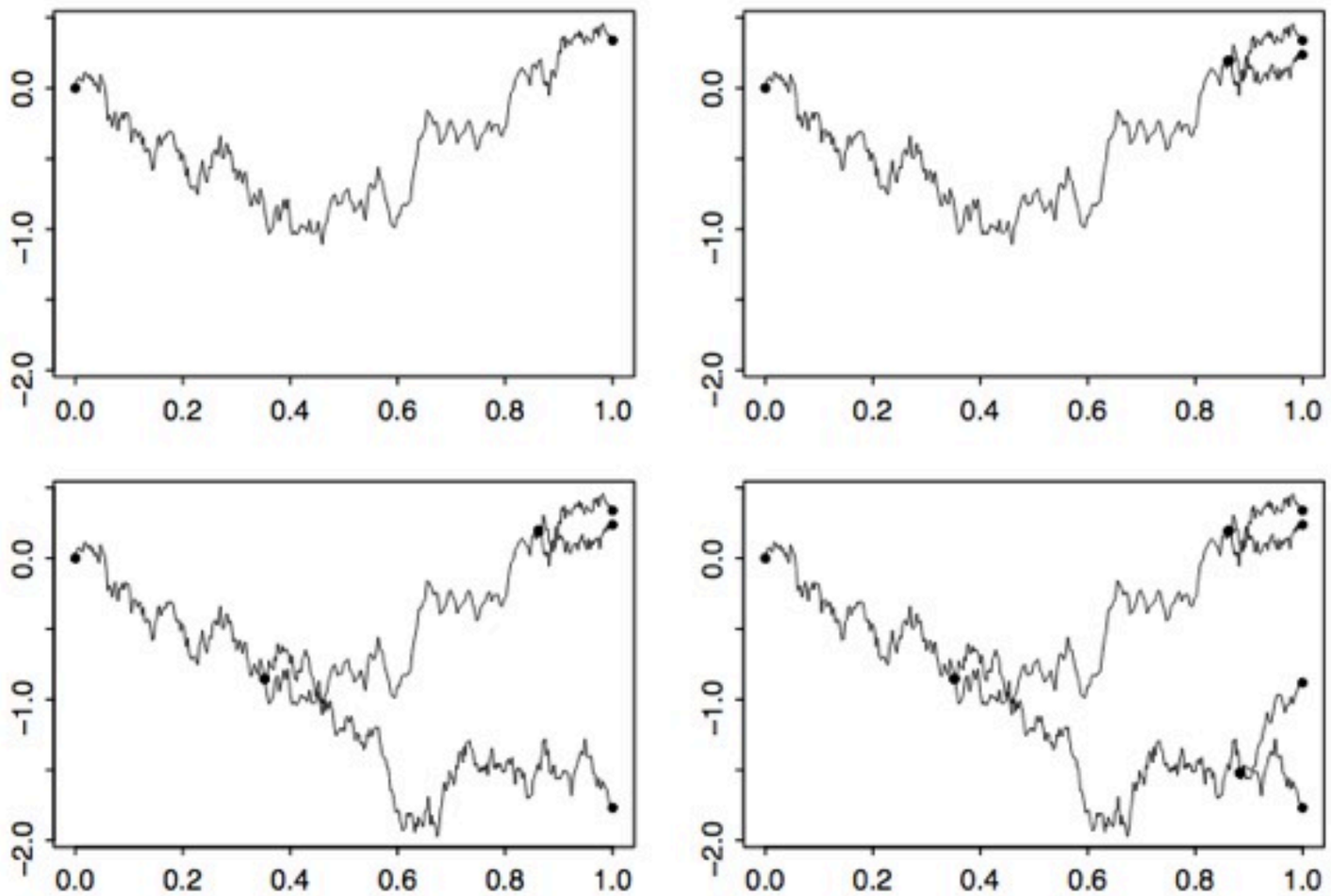
Continuum Limit of Partition-valued Markov Chains

Trees with Infinitely Many Levels

- Random trees described so far all consist of a finite number of levels L.
- We can be "nonparametric" about the number of levels of random trees.
- Allow a finite amount of change even with an infinite number of levels, by decreasing the change per level.



Dirichlet Diffusion Trees



In general, the ith point in the data set is obtained by following a path from the origin that initially coincides with the path to the previous $i-1$ data points. If the new path has not diverged at a time when paths to past data points diverged, the new path chooses between these past paths with probabilities proportional to the numbers of past paths that went each way. If at time t, the new path is following a path traversed by m previous paths, the probability that it will diverge from this path within an infinitesimal interval of duration dt is $a(t)dt/m$. Once divergence occurs, the new path moves independently of previous paths.

[Neal 2003]

Dirichlet Diffusion Trees

- The **Dirichlet diffusion tree** (DFT) hierarchical partitioning structure can be derived from the continuum limit of a nCRP:
  - Start with the null partition $\varrho_0 = \{[n]\}$.
  - For each time t, define

$$\varrho_{t+dt} = \mathrm{FRAG}(\varrho_t, 0, a(t)dt)$$

- The continuum limit of the Markov chain of partitions becomes a *continuous time partition-valued Markov process*: a **fragmentation process**.
- Generalization to **Pitman-Yor diffusion trees**.

[Knowles & Ghahramani 2011]

Kingman's Coalescent

- Taking the continuum limit of the one-parameter (Markov chain) CRF leads to another partition-valued Markov process: **Kingman's coalescent**.
  - Start with the trivial partition $\varrho_0 = \{\{1\},\{2\},...,\{n\}\}$.
  - For each time $t < 0$:

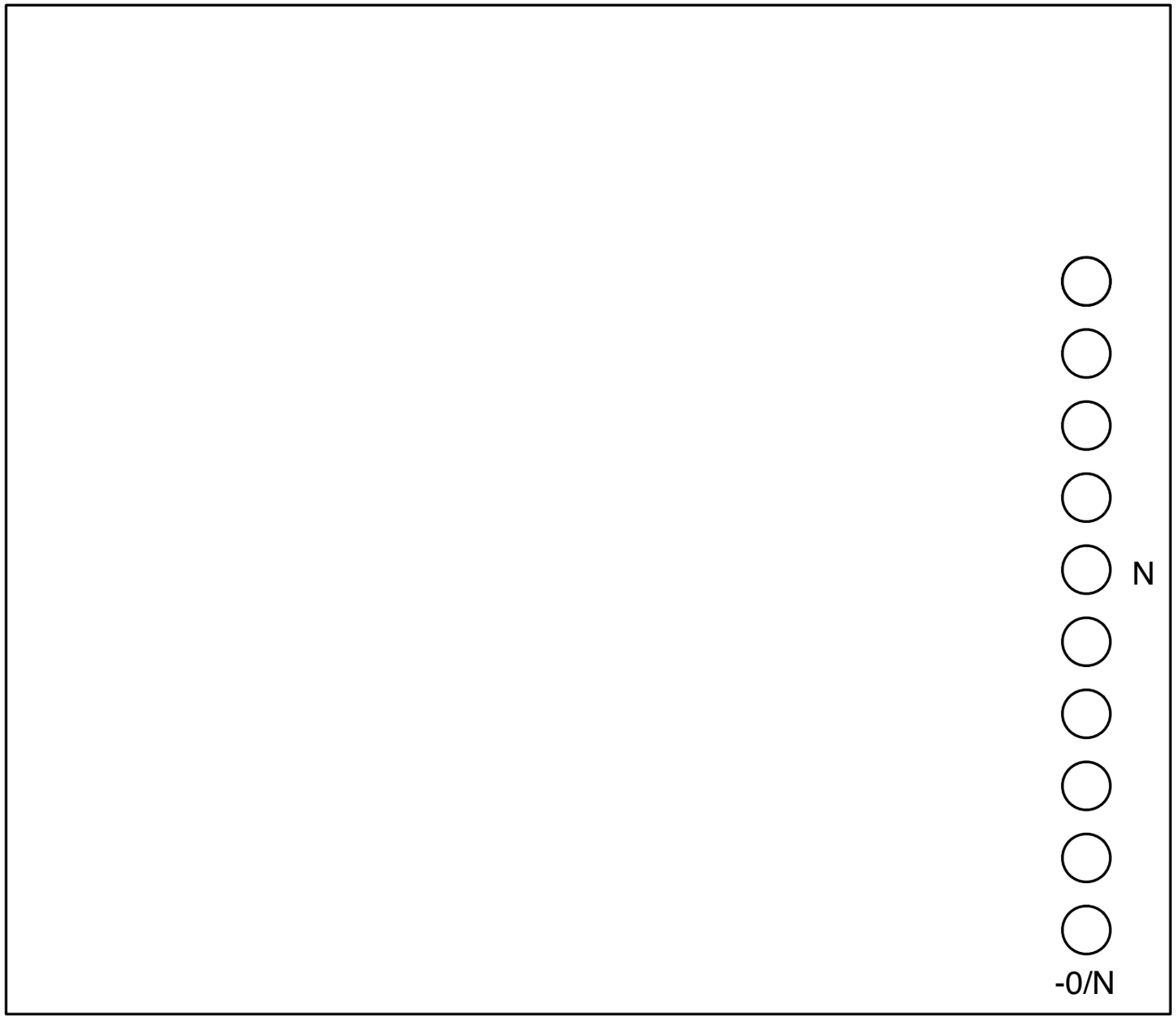
$$\varrho_{t-dt} = \mathrm{COAG}(\varrho_t, 0, a(t)/dt)$$

    - This is the simplest example of a **coalescence or coagulation process**.
- A standard genealogical process in genetics.
- A generalization called **Λ-coalescent**.

[Kingman 1982a,b, Pitman 1999]

Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.

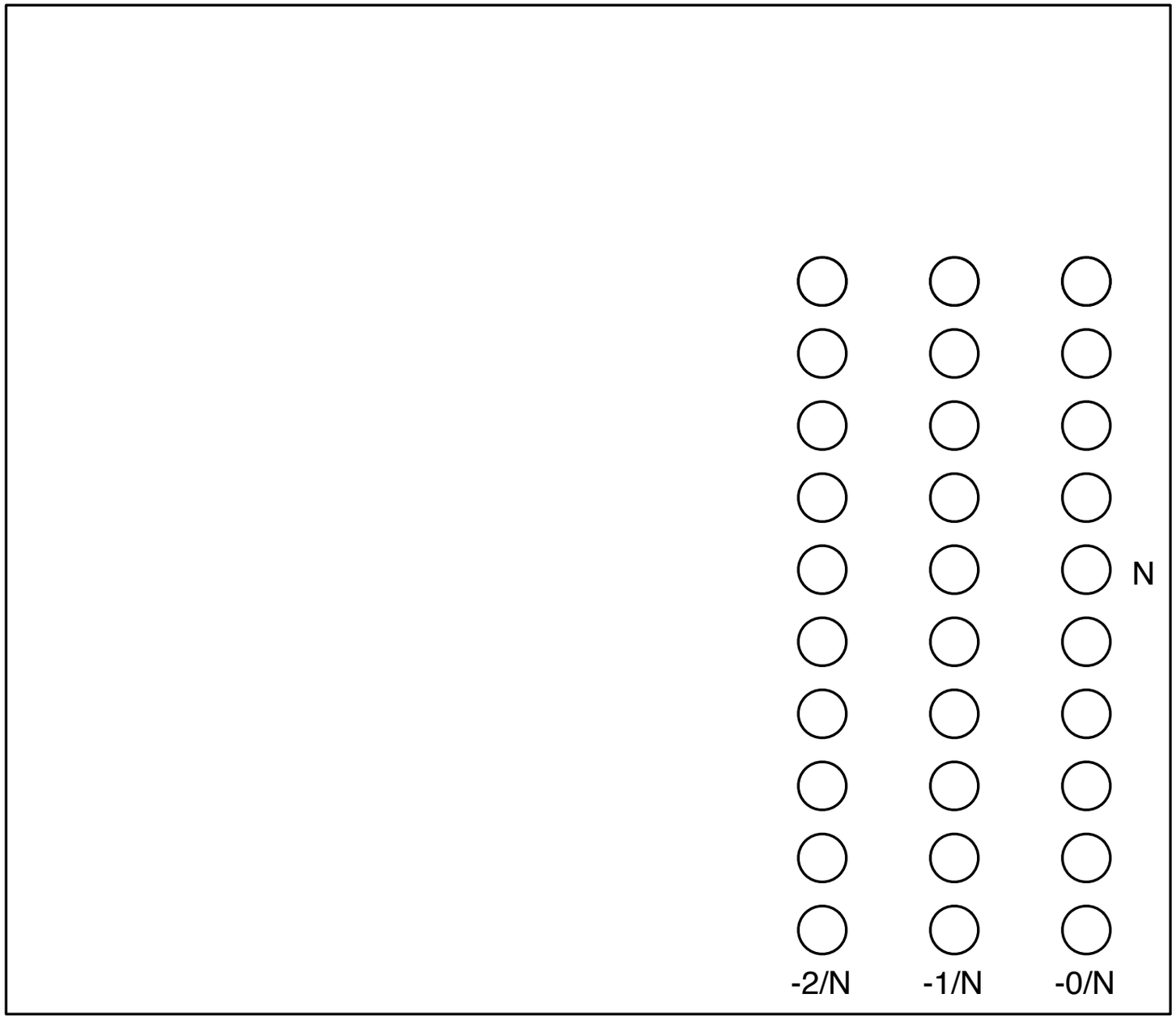


Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.

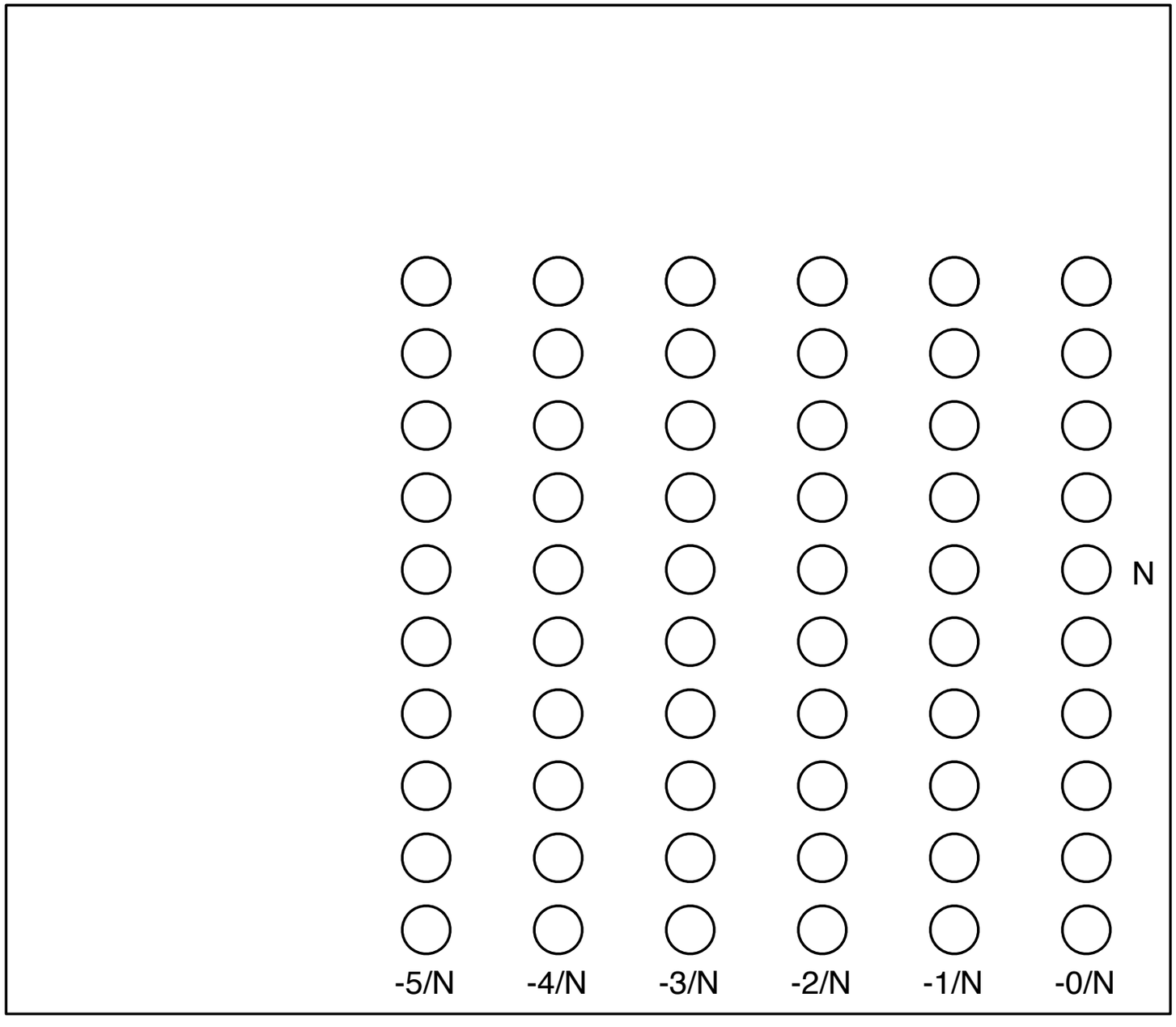
Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.



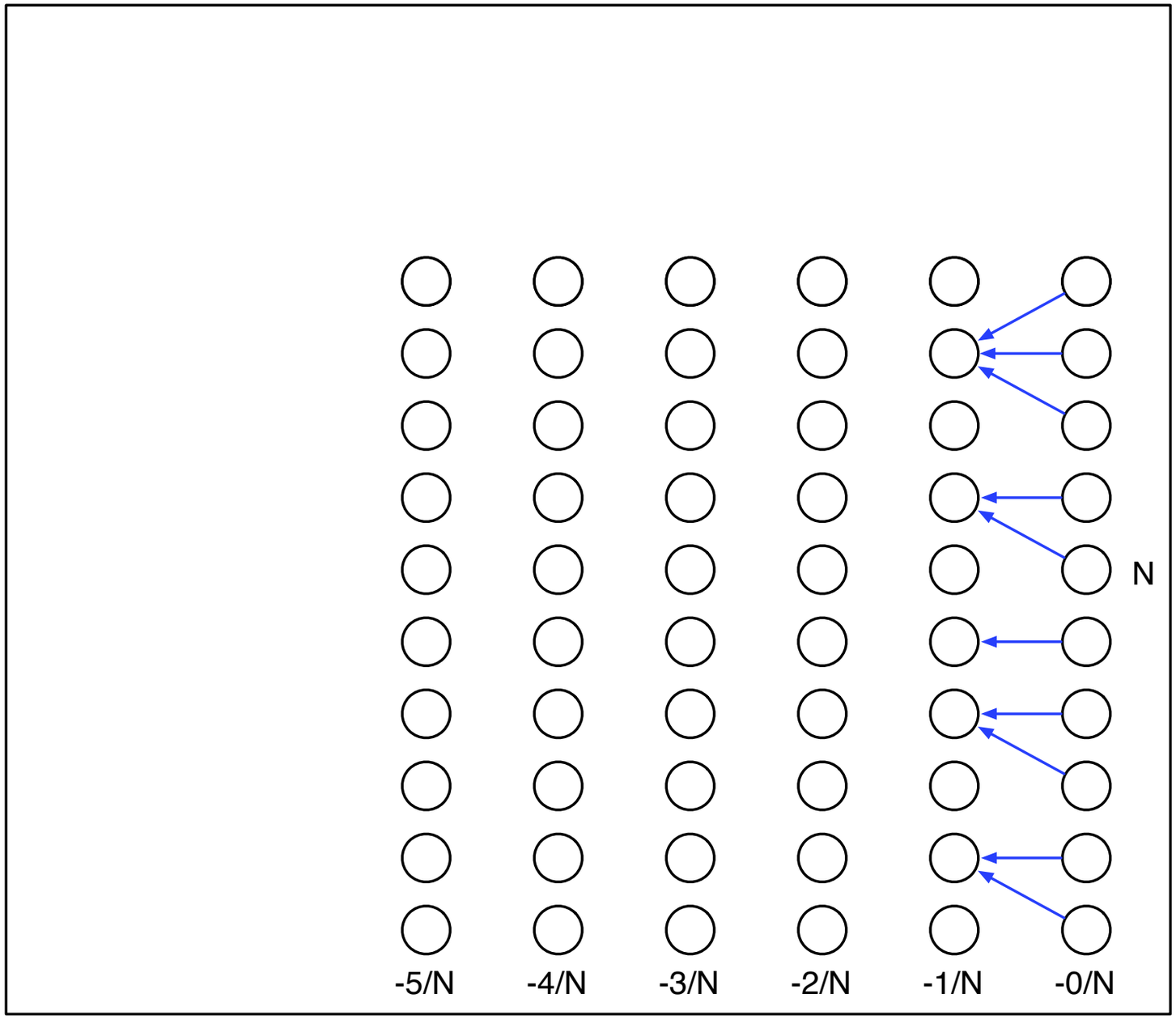
Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.



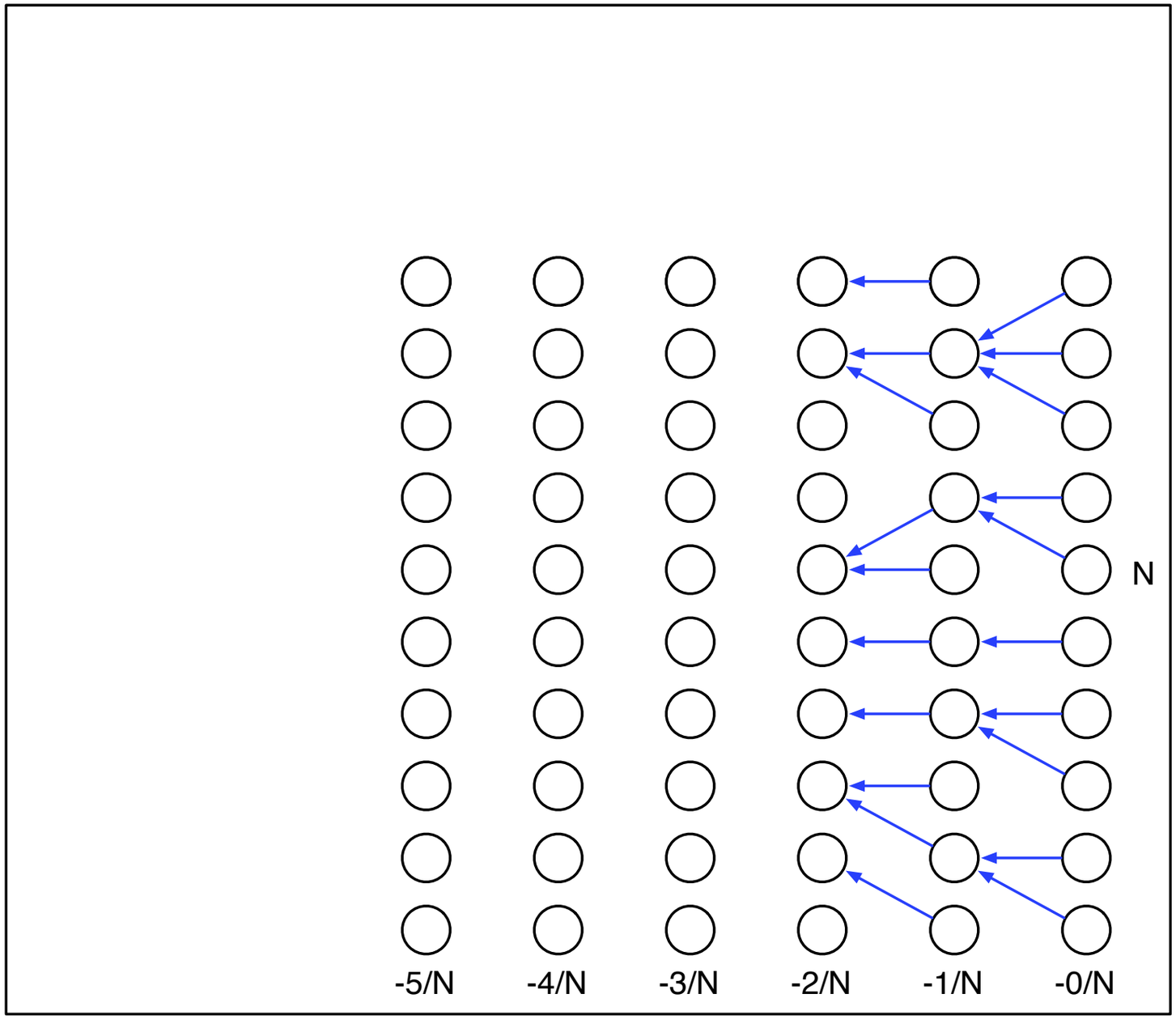
Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.



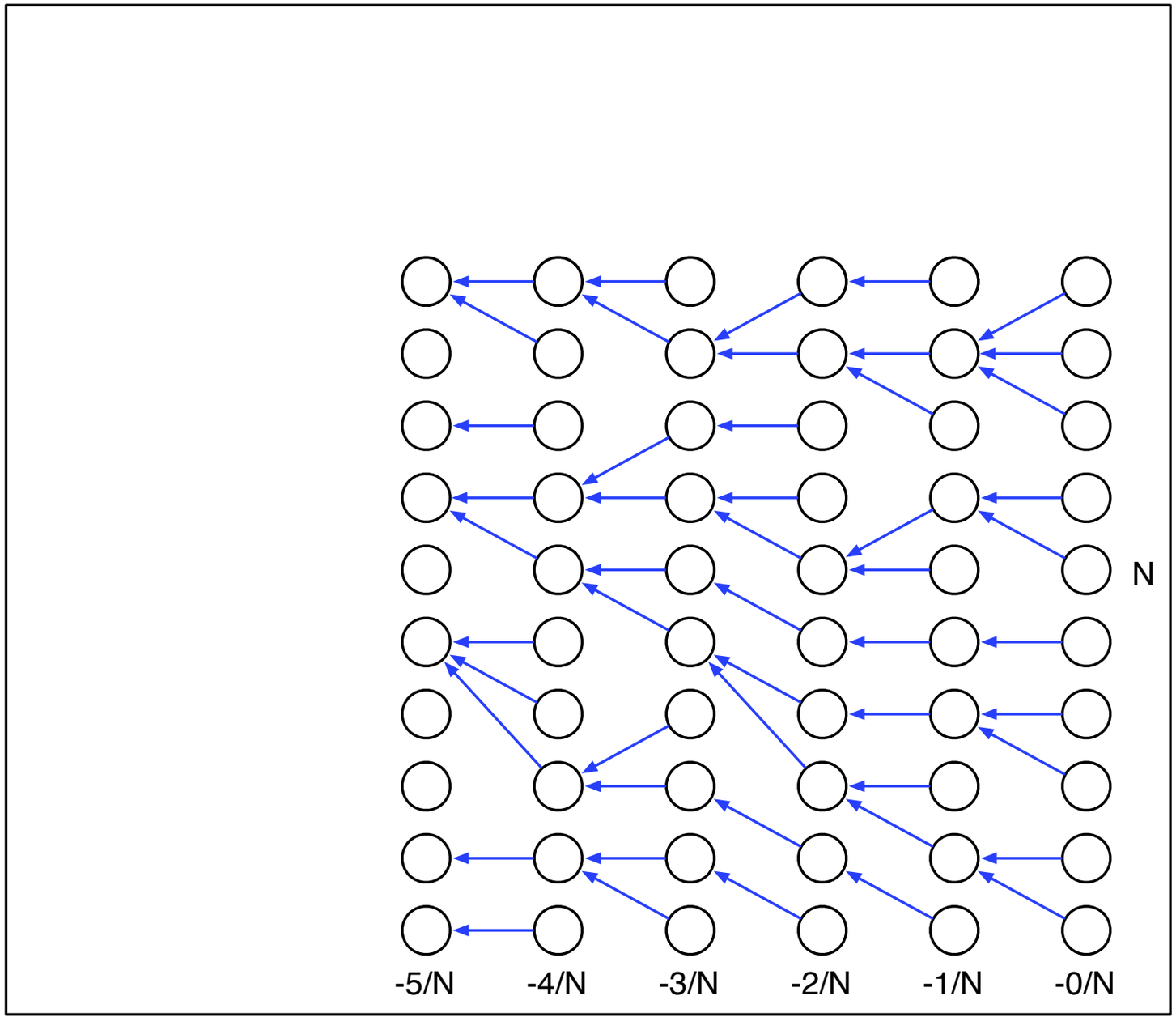
Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.



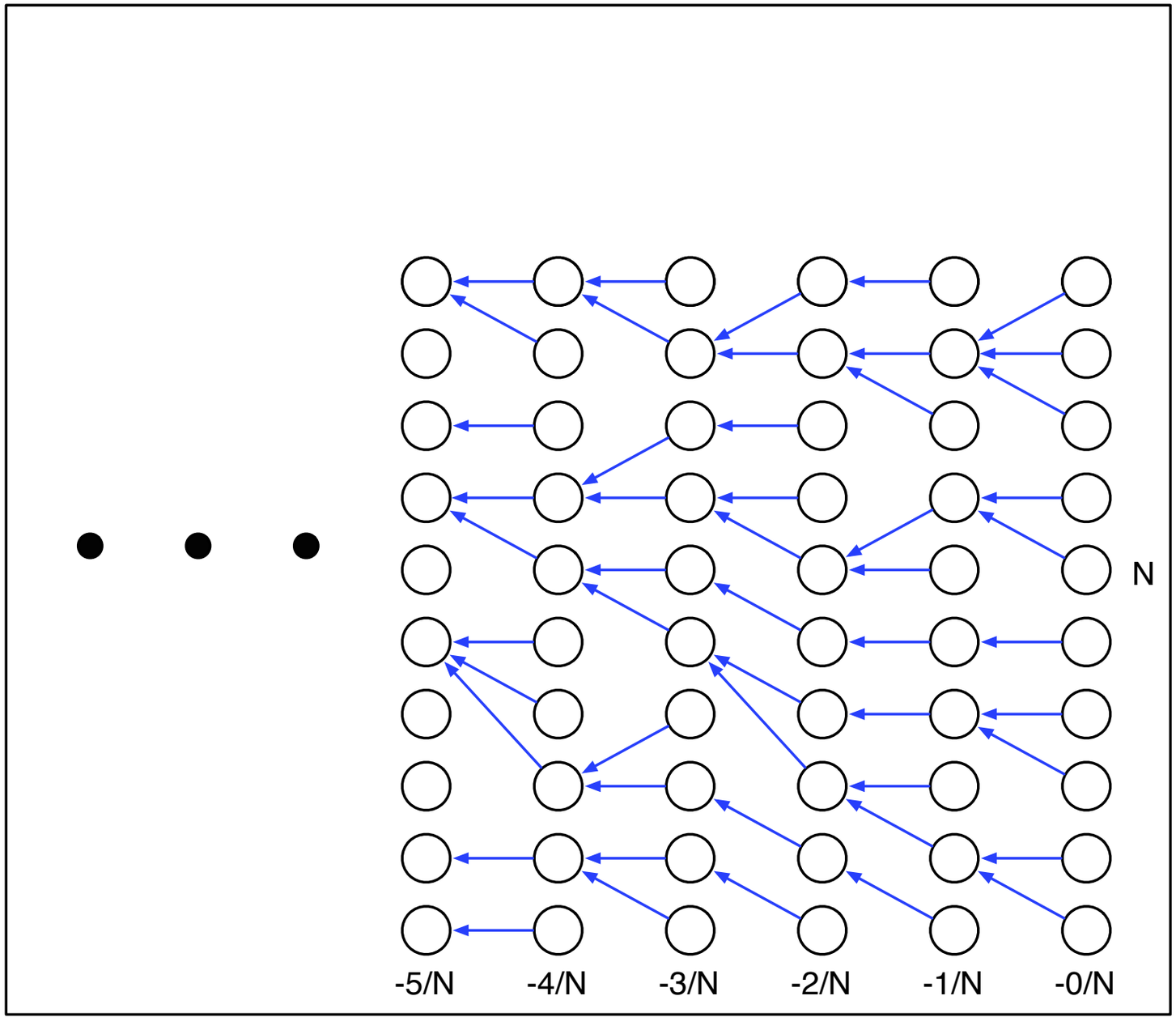
Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.



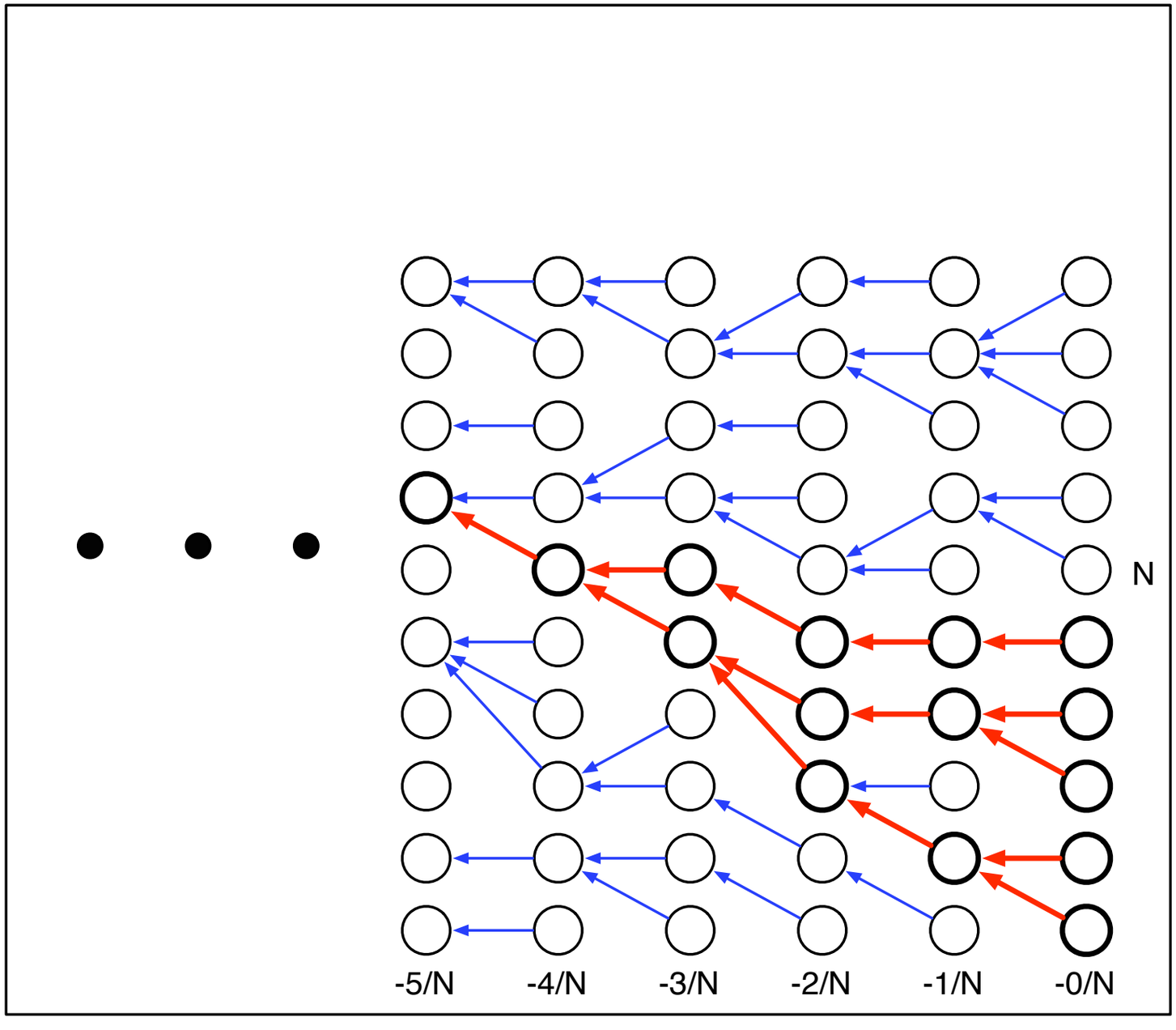
Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.



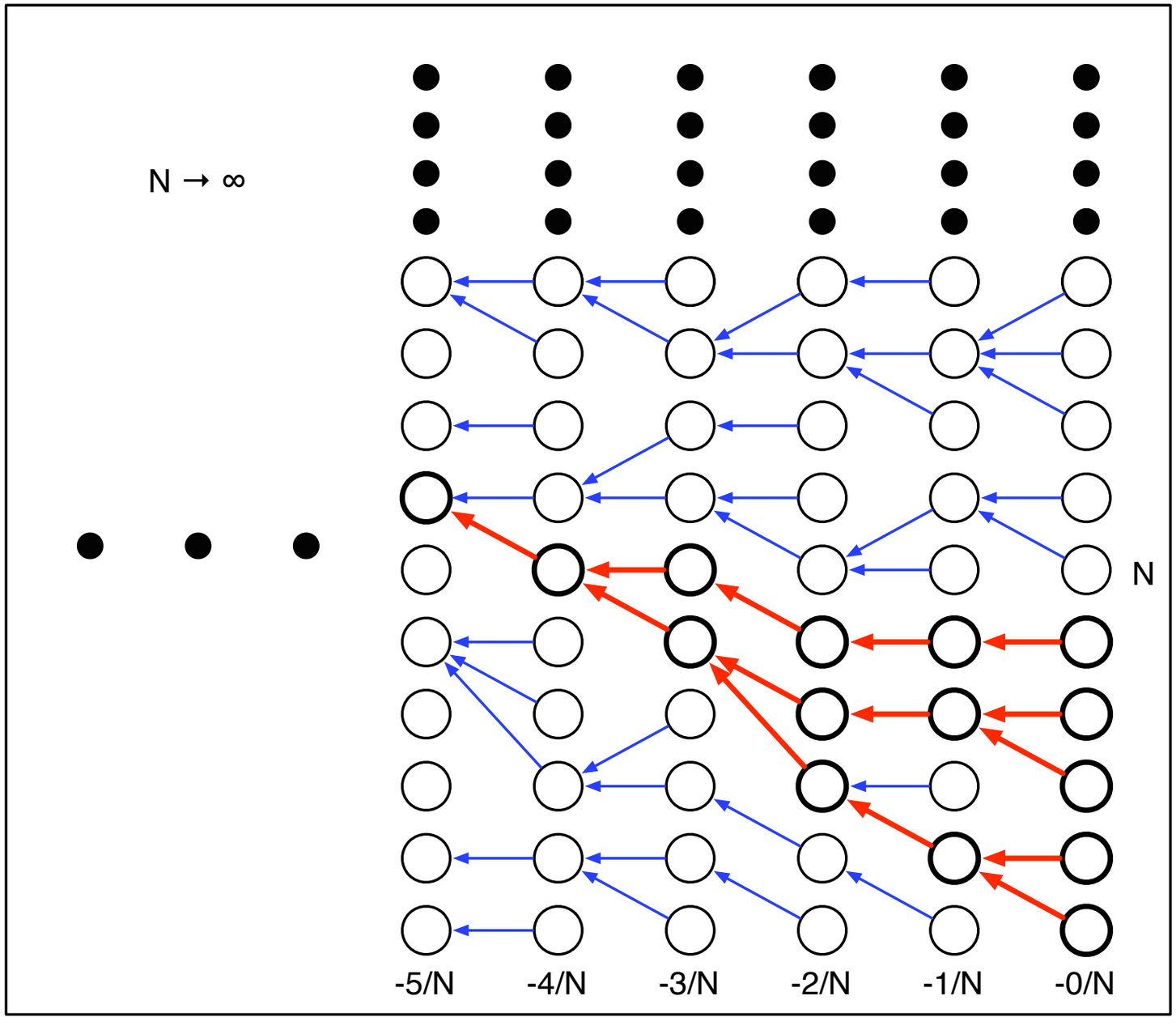
Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.

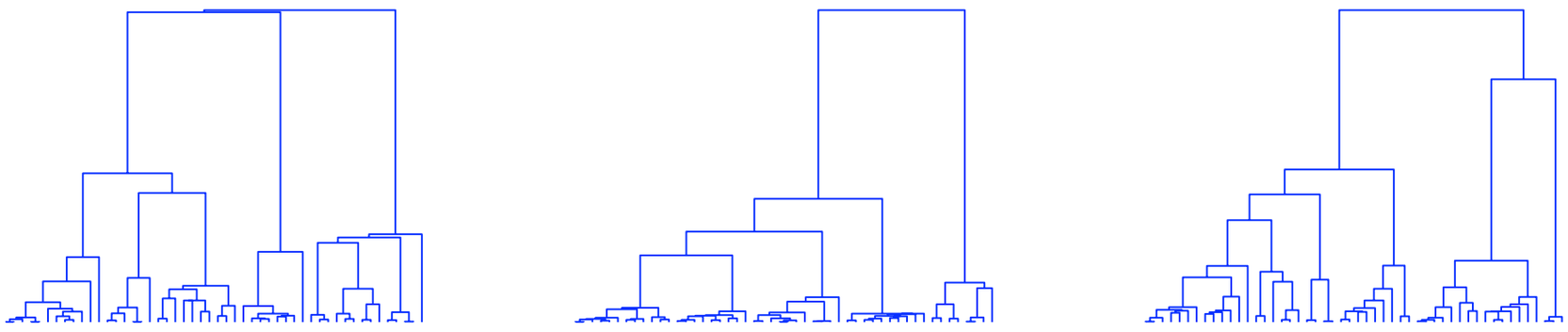


Kingman's Coalescent

- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.



Kingman's Coalescent



Binary Ultrametric Random Trees

- Both Dirichlet diffusion trees and Kingman's coalescent are priors over binary trees, i.e. every internal node has exactly 2 children.
  - Generalizations allow for more than 2 children.
- Both models are priors over ultrametric trees, i.e. all observations are at leaves which are equidistant from the root.
  - Can generalize by allowing observations at different distances from root.

- Constructions for other types of random trees:
  - Gibbs fragmentation trees
  - Continuum random trees
  - Standard additive coalescent

Sequence Memoizer

Markov Models for Language and Text

- Usually makes a Markov assumption to simplify model:

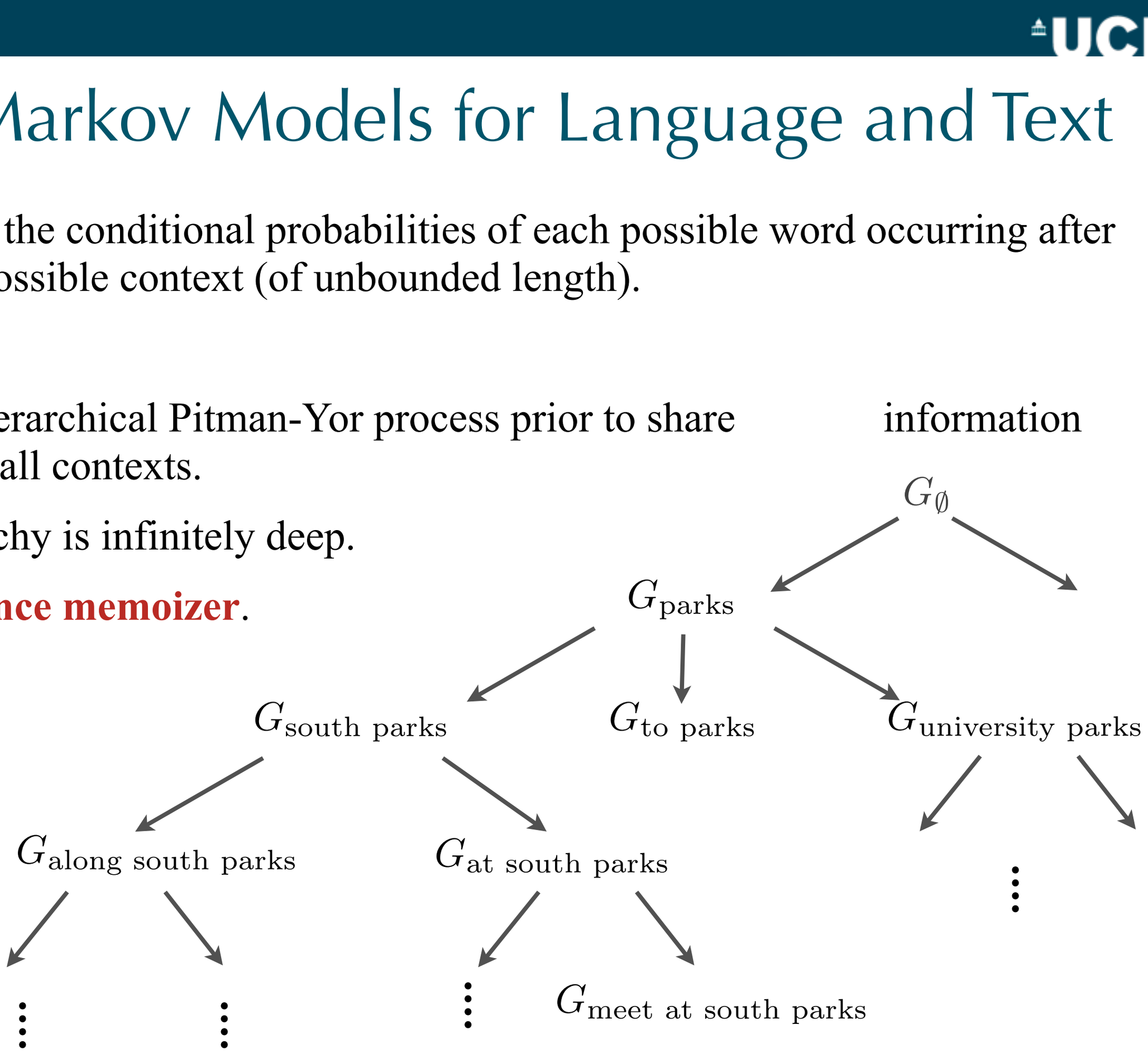
P(south parks road) ~
P(south)*
P(parks | south)*
P(road | south parks)

- Language models: usually Markov models of order 2-4 (3-5-grams).
- How do we determine the order of our Markov models?
- Is the Markov assumption a reasonable assumption?
  - Be nonparametric about Markov order...

Non-Markov Models for Language and Text

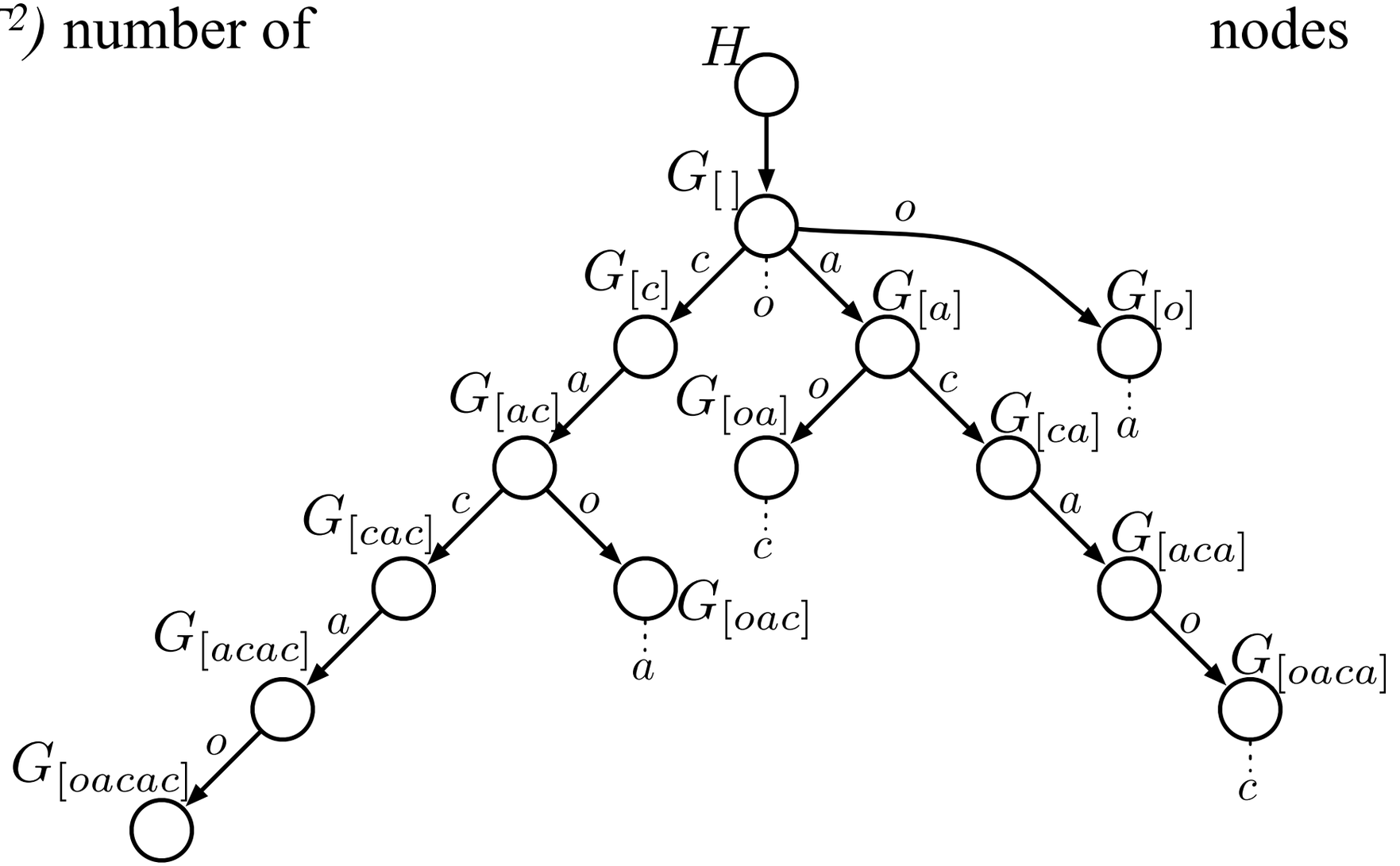
- Model the conditional probabilities of each possible word occurring after each possible context (of unbounded length).

- Use hierarchical Pitman-Yor process prior to share information across all contexts.
- Hierarchy is infinitely deep.
- **Sequence memoizer**.



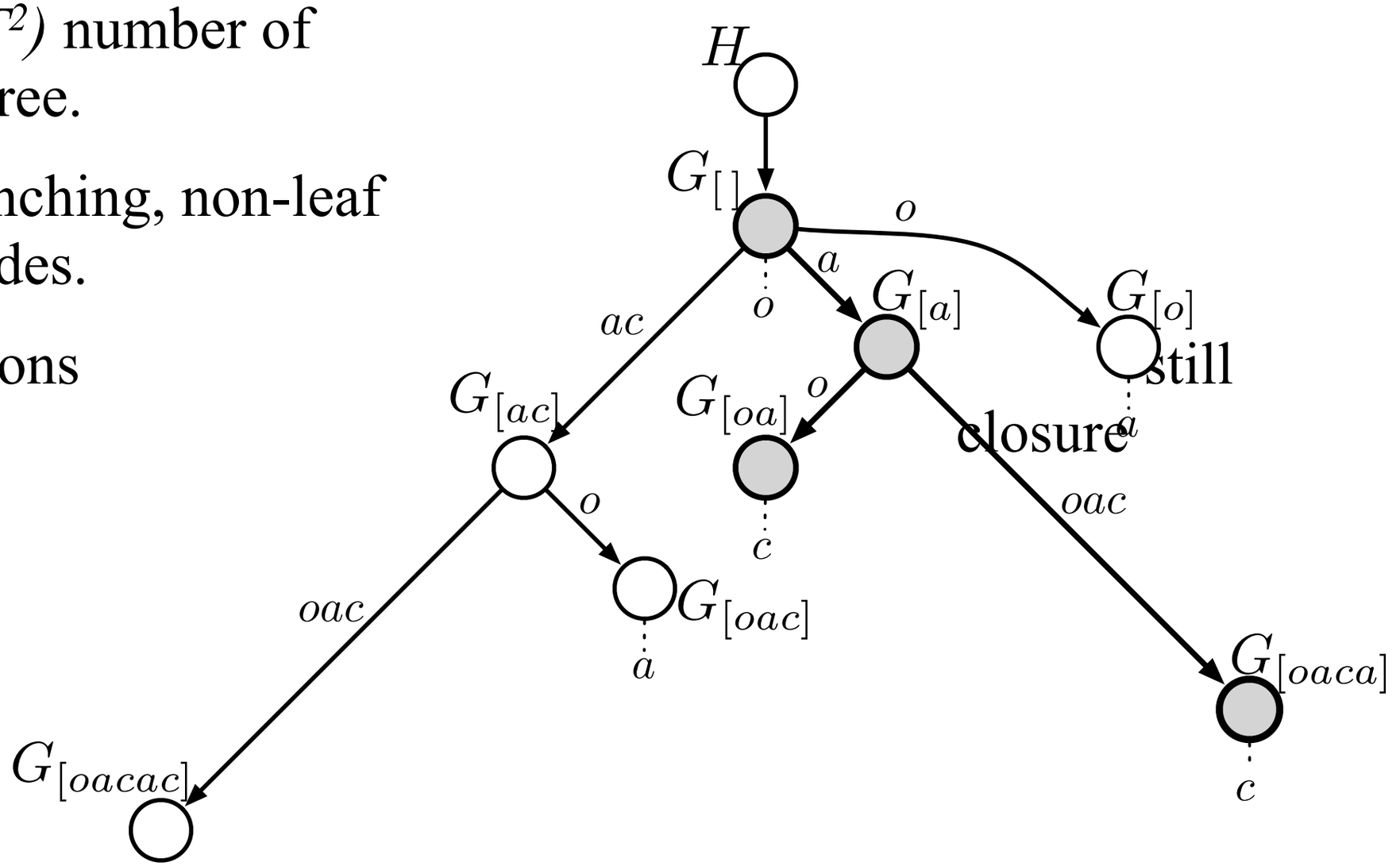
Model Size: Infinite -> $O(T^2)$

- The sequence memoizer model is very large (actually, infinite).
- Given a training sequence (e.g.: o,a,c,a,c), most of the model can be ignored (integrated out), leaving a finite number of nodes in context tree.
- But there are still $O(T^2)$ number of nodes in the context tree.



Model Size: Infinite -> $O(T^2)$ -> $O(2T)$

- The sequence memoizer model is very large (actually, infinite).
- Given a training sequence (e.g.: o,a,c,a,c), most of the model can be ignored (integrated out), leaving a finite number of nodes in context tree.
- But there are still $O(T^2)$ number of nodes in the context tree.
- Integrate out non-branching, non-leaf nodes leaves $O(T)$ nodes.
- Conditional distributions Pitman-Yor due to property.



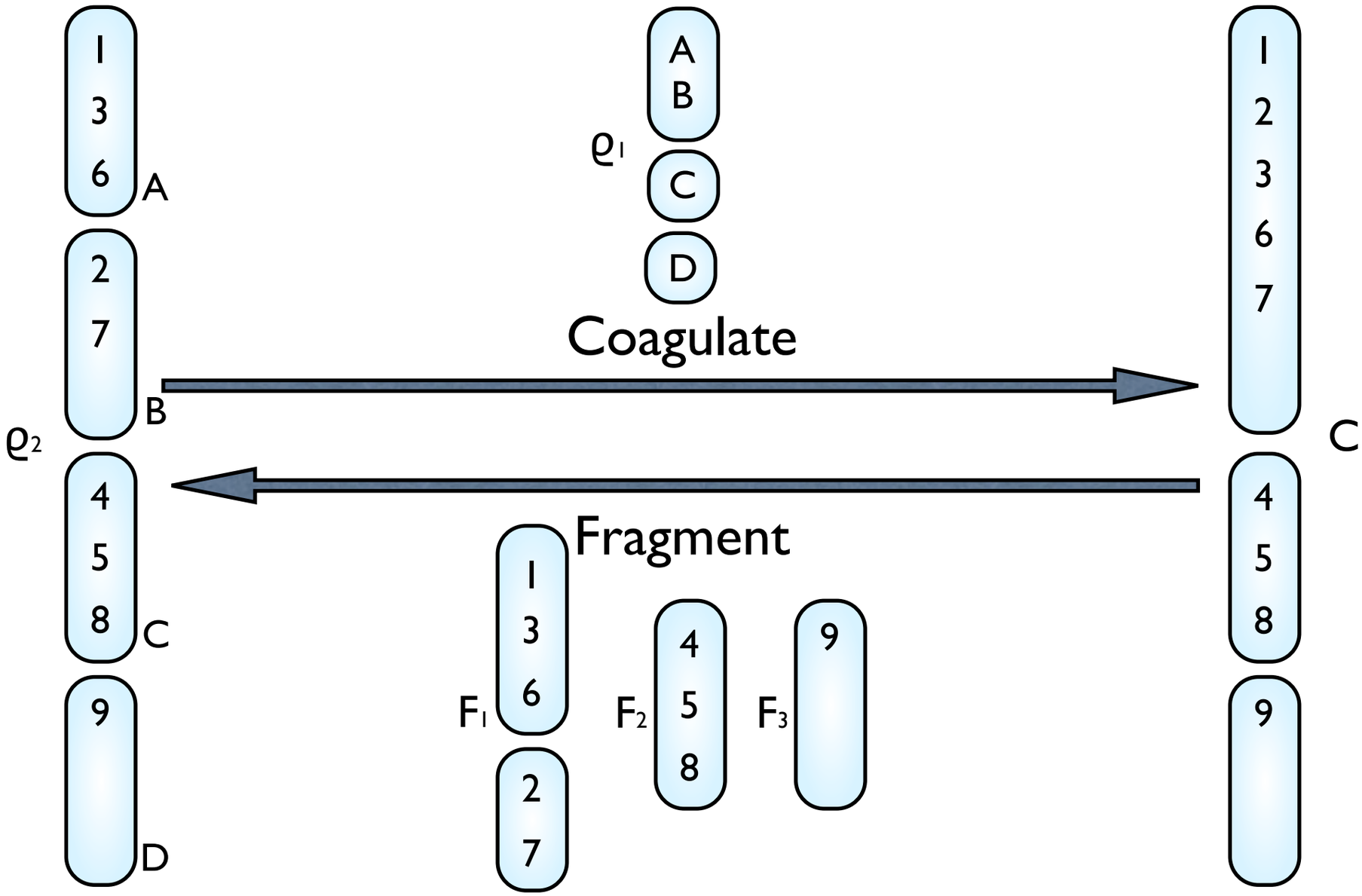
closure still

Duality of Coagulation and Fragmentation

- The following statements are equivalent:

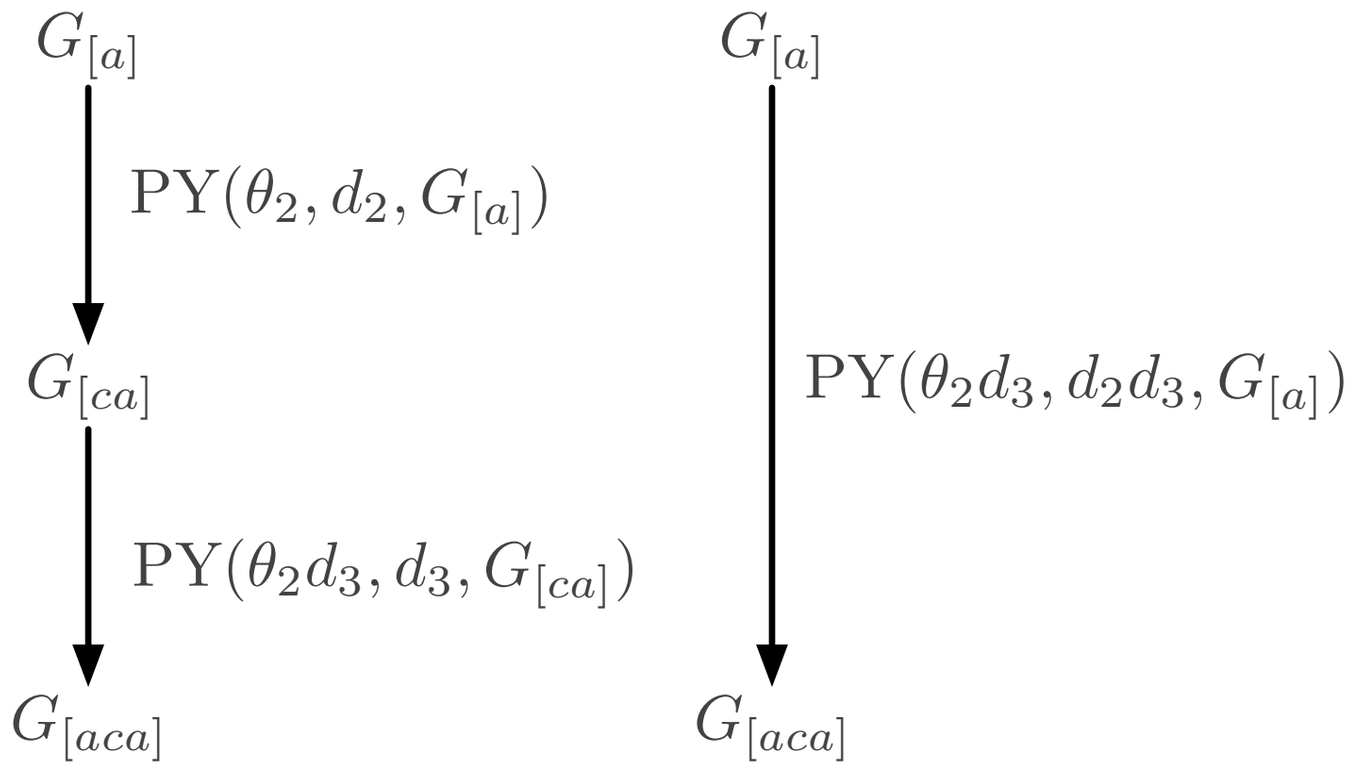
(I) $\varrho_2 \sim \mathrm{CRP}([n], d_2, \alpha d_2)$ and $\varrho_1 | \varrho_2 \sim \mathrm{CRP}(\varrho_2, d_1, \alpha)$

(II) $C \sim \mathrm{CRP}([n], d_1 d_2, \alpha d_2)$ and $F_c | C \sim \mathrm{CRP}(c, d_2, -d_1 d_2) \quad \forall c \in C$



Closure under Marginalization

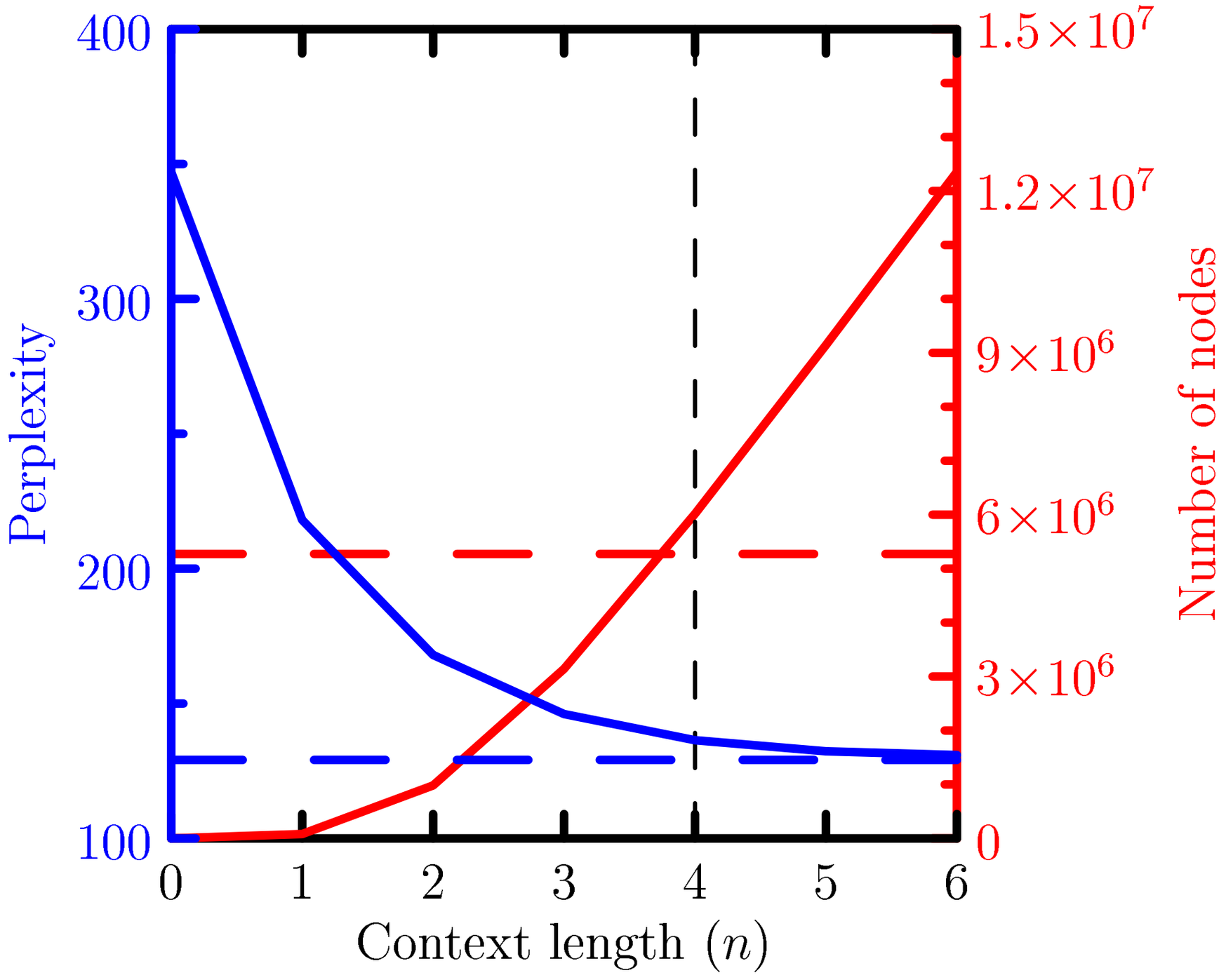
- Marginalizing out internal Pitman-Yor processes is equivalent to coagulating the corresponding Chinese restaurant processes.



- Fragmentation and coagulation duality means that the coagulated partition is also Chinese restaurant process distributed.
- Corresponding Pitman-Yor process is the resulting marginal distribution of $G_{[aca]}$.

[Wood et al 2009, Gasthaus & Teh 2010]

Comparison to Finite Order HPYLM



Compression Results

Model	Average bits/byte
gzip	2.61
bzip2	2.11
CTW	1.99
PPM	1.93
Sequence Memoizer	1.89

Calgary corpus
SM inference: particle filter
PPM: Prediction by Partial Matching
CTW: Context Tree Weigting
Online inference, entropic coding.

Fragmentation-Coagulation Processes

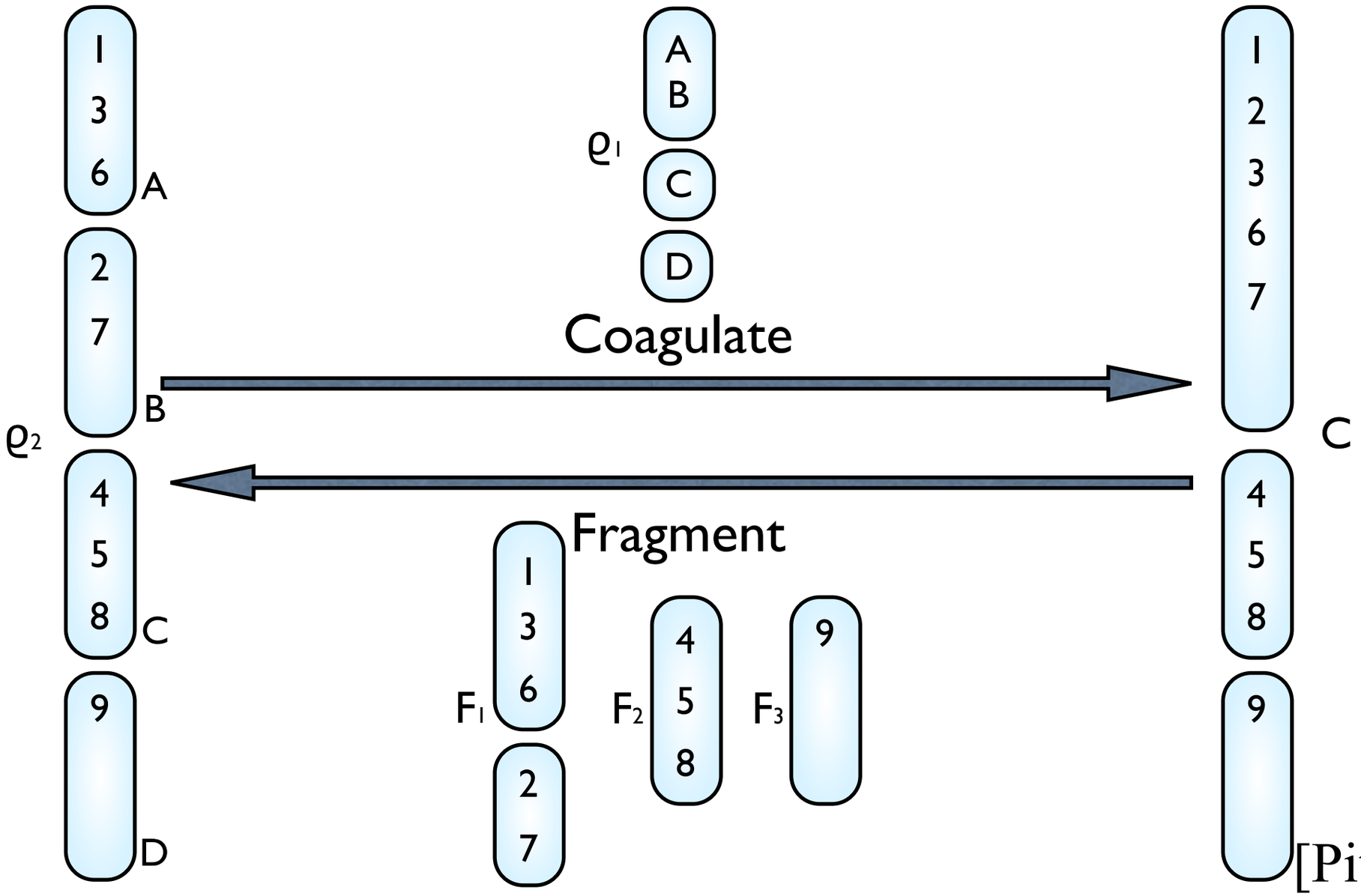
[Berestycki 2004, Teh et al 2011]

Duality of Coagulation and Fragmentation

- The following statements are equivalent:

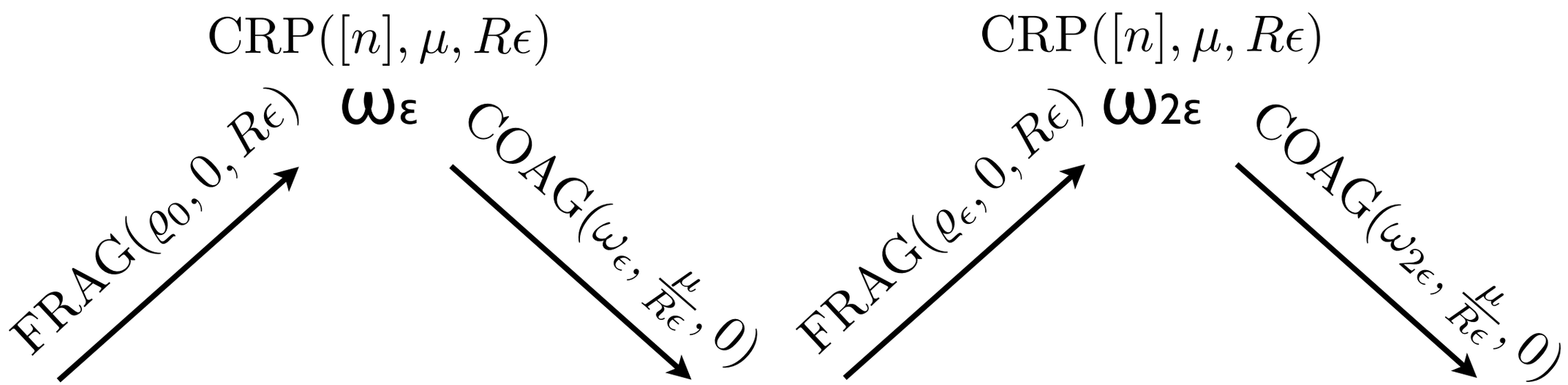
(I) $\varrho_2 \sim \mathrm{CRP}([n], d_2, \alpha d_2)$ and $\varrho_1 | \varrho_2 \sim \mathrm{CRP}(\varrho_2, d_1, \alpha)$

(II) $C \sim \mathrm{CRP}([n], d_1 d_2, \alpha d_2)$ and $F_c | C \sim \mathrm{CRP}(c, d_2, -d_1 d_2) \quad \forall c \in C$



[Pitman 1999]

Markov Chain over Partitions

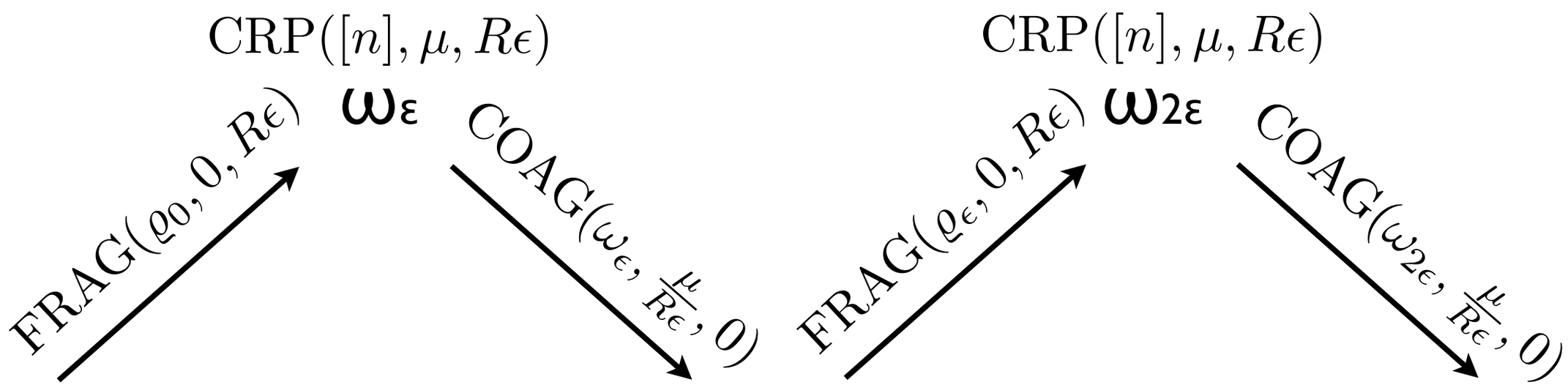


ϱ_0 ϱ_ϵ $\varrho_{2\epsilon}$

$\mathrm{CRP}([n], \mu, 0)$ $\mathrm{CRP}([n], \mu, 0)$ $\mathrm{CRP}([n], \mu, 0)$

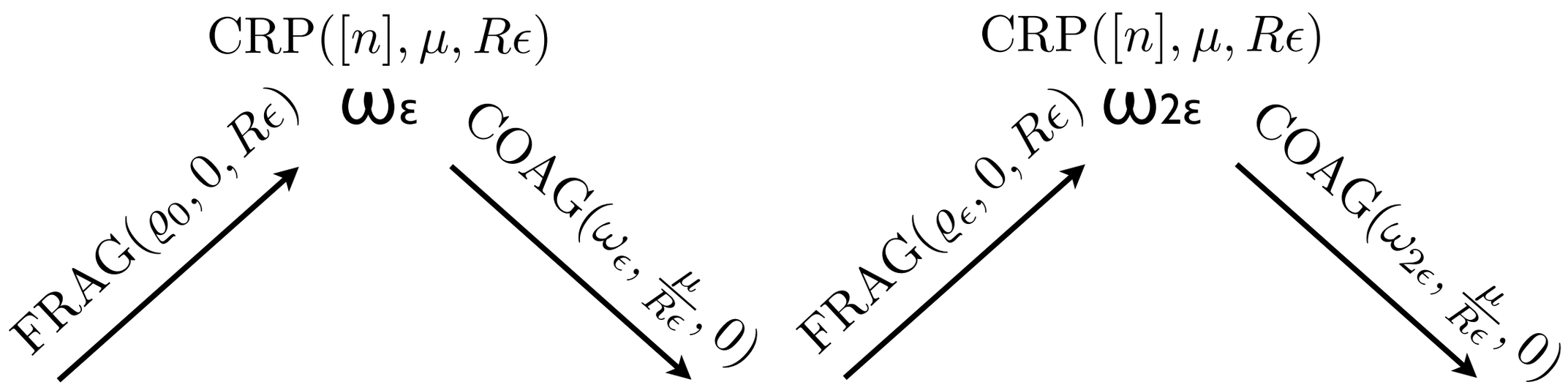
- Defines a Markov chain over partitions.
- Each transition is a fragmentation followed by coagulation.

Stationary Distribution



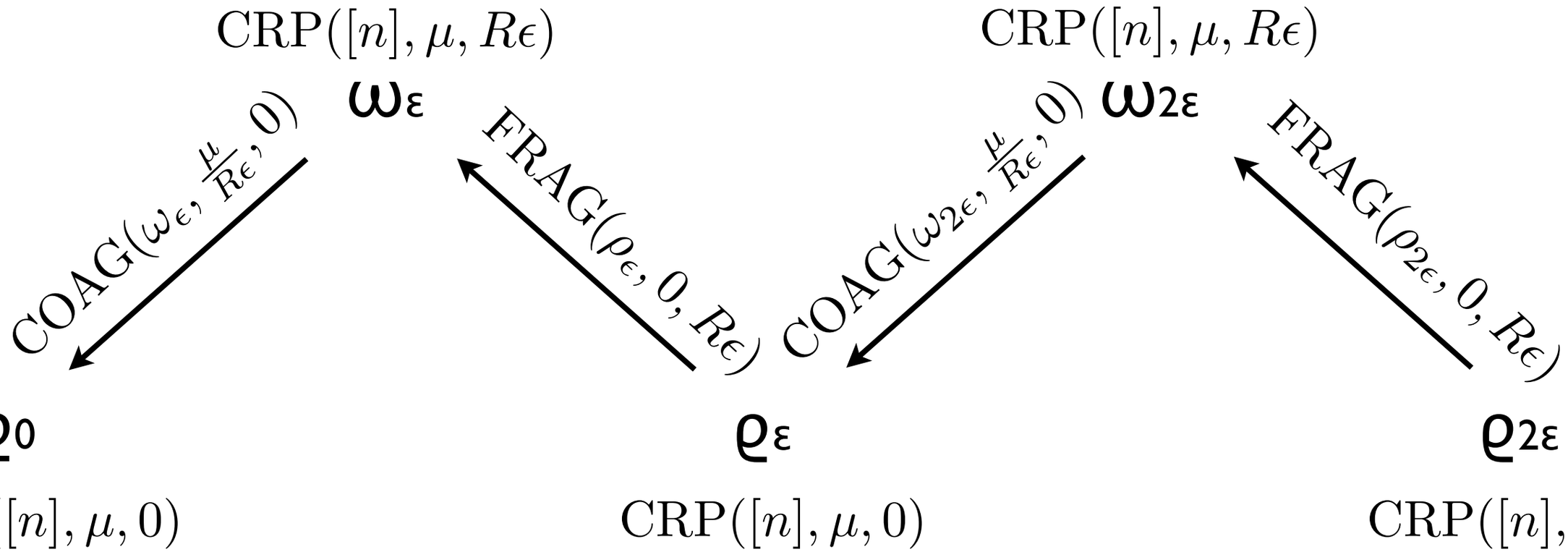
- *Stationary distribution* is a CRP with parameters μ and 0.

Exchangeability and Projectivity



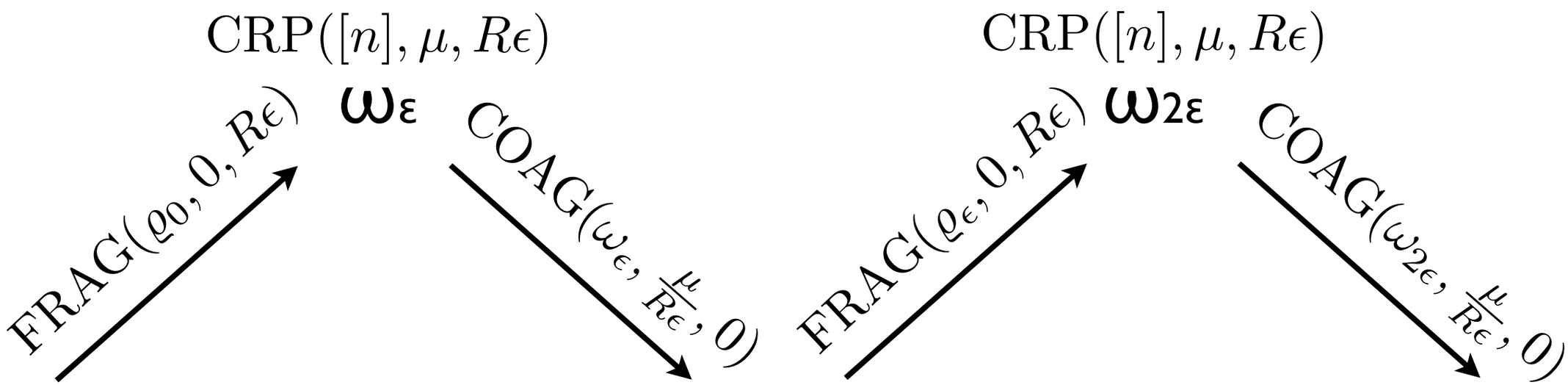
- Each π_t is exchangeable, so that the whole Markov chain is an *exchangeable process*.
- Projectivity of the Chinese restaurant process extends to the Markov chain as well.

Reversibility of Markov Chain



- The Markov chain is reversible.
- Coagulation and fragmentation are duals of each other.

Continuum Limit

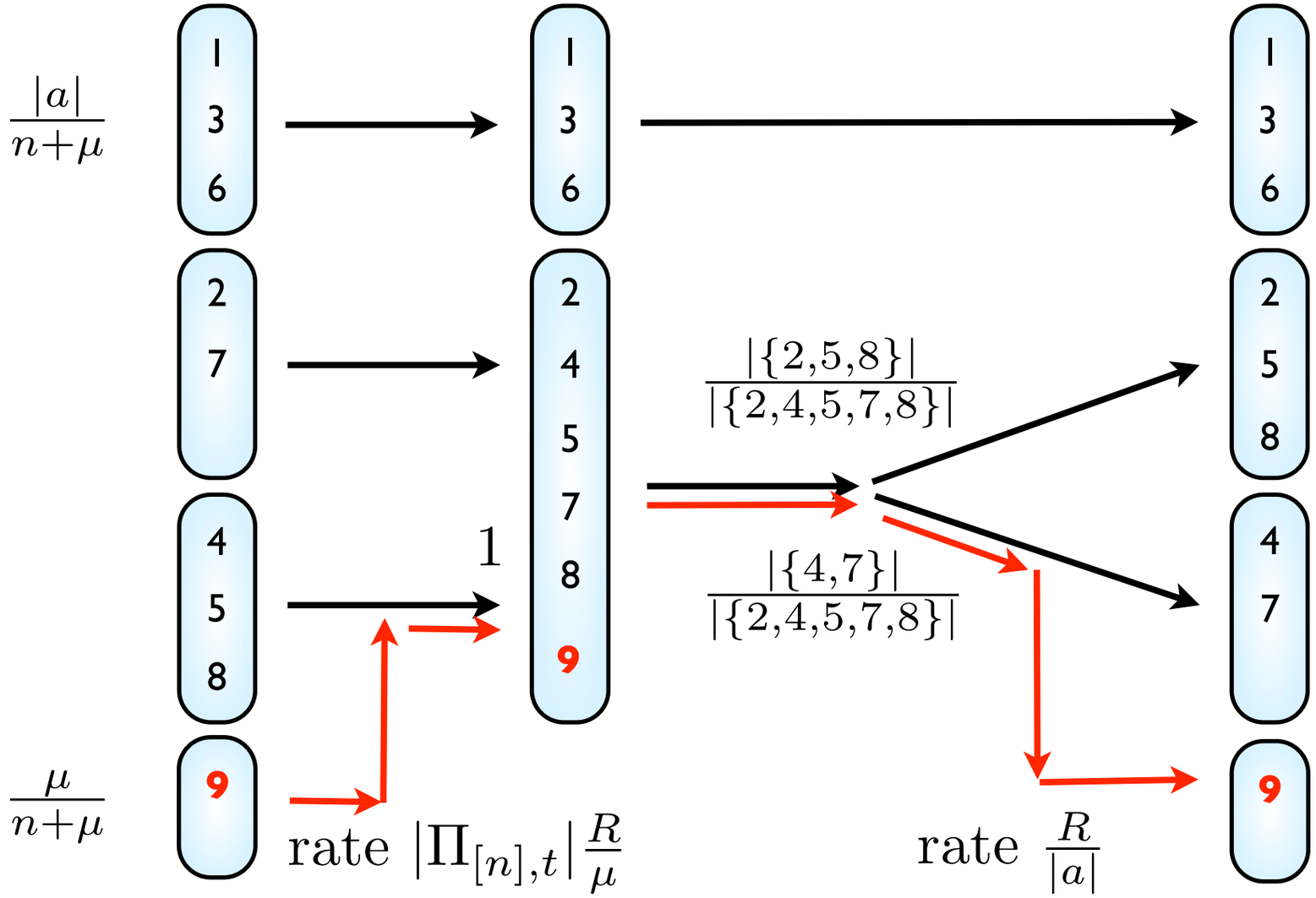


ϱ_0 ϱ_ε $\varrho_{2\varepsilon}$

$\mathrm{CRP}([n], \mu, 0)$ $\mathrm{CRP}([n], \mu, 0)$ $\mathrm{CRP}([n], \mu, 0)$

- Taking $\varepsilon \to 0$ obtains a continuous time Markov process over partitions, an **exchangeable fragmentation-coalescence process** (Berestycki 2004).
- At each time, at most one coagulation (involving two blocks) or one fragmentation (splitting into two blocks) will occur.

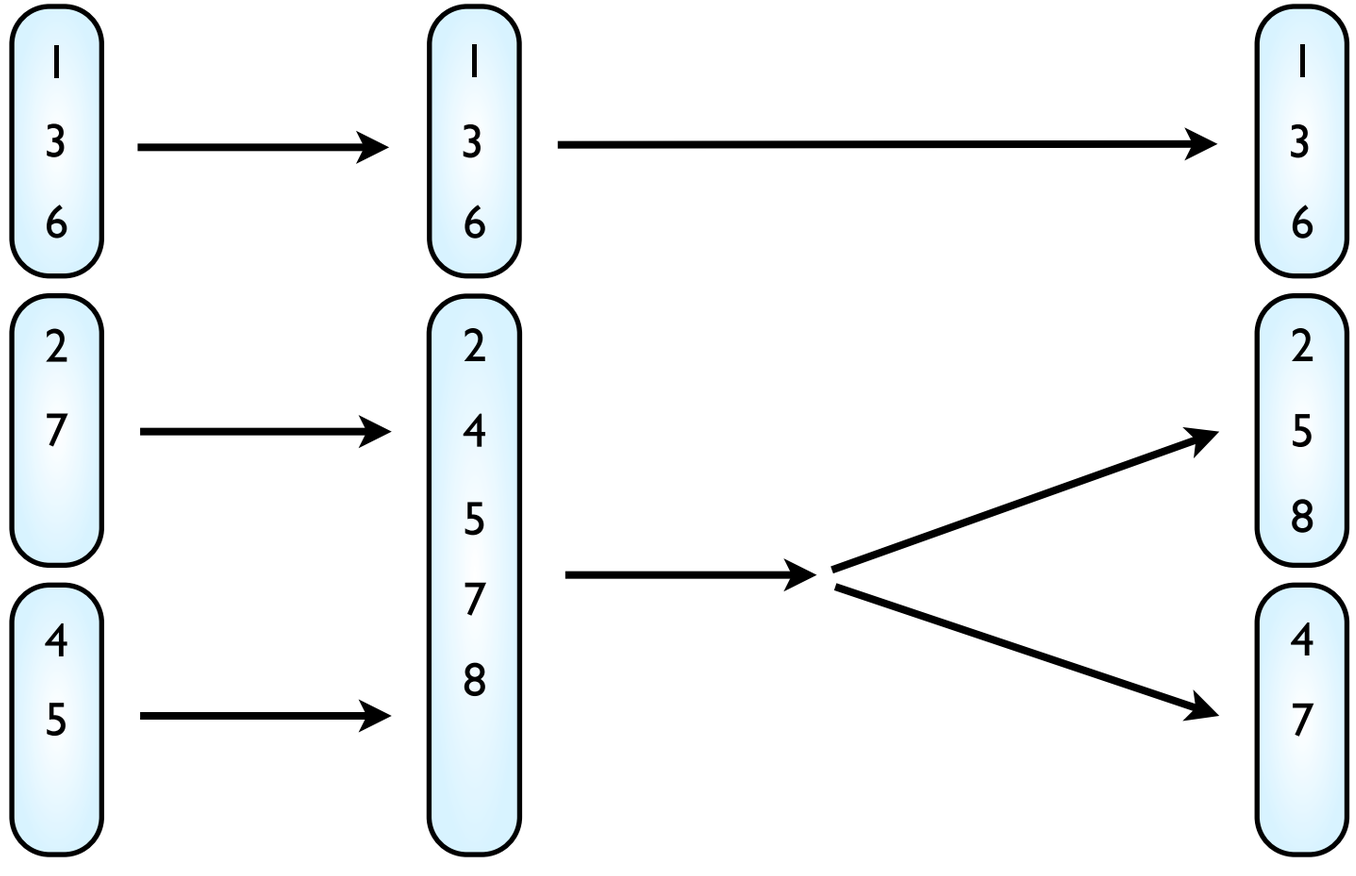
Conditional Distribution of a Trajectory



- This process is reversible.

Coagulation and Fragmentation Rates

- Describe Markov process in terms of rates of fragmentation and coagulation events:
  - Rate of fragmentation of a $\in \Pi_{[n],t}$ into b and c: $R\frac{\Gamma(|b|)\Gamma(|c|)}{\Gamma(|a|)}$
  - Rate of coagulation of a,b $\in \Pi_{[n],t}$ into into a$\bigcup$b: R/μ



Number of Clusters and Events

- Over an interval T, if Π_t are partitions of [n], then the expected number clusters at each point t is:

$$\mu(\psi(n+\mu) - \psi(\mu)) = O(\mu \log(n - \tfrac{1}{2} + \mu))$$

- Expected number of fragmentation and coagulation events is:

$$R\mu T \left((\psi(n+\mu) - \psi(\mu))^2 + \psi'(n+\mu) - \psi'(\mu) \right)$$

$$= O\left(R\mu T \log(n - \tfrac{1}{2} + \mu)^2 \right)$$

Dirichlet Diffusion Trees and Coalescents

- Rate of fragmentation is same as for Dirichlet diffusion trees with constant fragmentation rate (Neal 2003).
- Rate of coagulation is same as for the coalescent (with time rescaled) (Kingman 1982).
- Reversibility means that the Dirichlet diffusion tree is the “reverse” of Kingman’s coalescent.
- Class of exchangeable fragmentation-coalescence processes (Berestycki 2004) includes more general processes.
    - This process seems to be a canonical example of exchangeable fragmentation-coalescence processes, but cannot find a reference in literature?

Relationship with Hidden Markov Models

- Both can be interpreted as models of sequence data with a latent partition structure at each time point.
- Hidden Markov models have explicit labels of hidden states, fragmentation-coagulation processes do not.
- Hidden Markov models need to specify the number of states, fragmentation-coagulation processes do not.
- HMM labels allow generalization across times, but lead to label switching problems.

Comparison with Bayesian HMMs

data:

0000000000000000

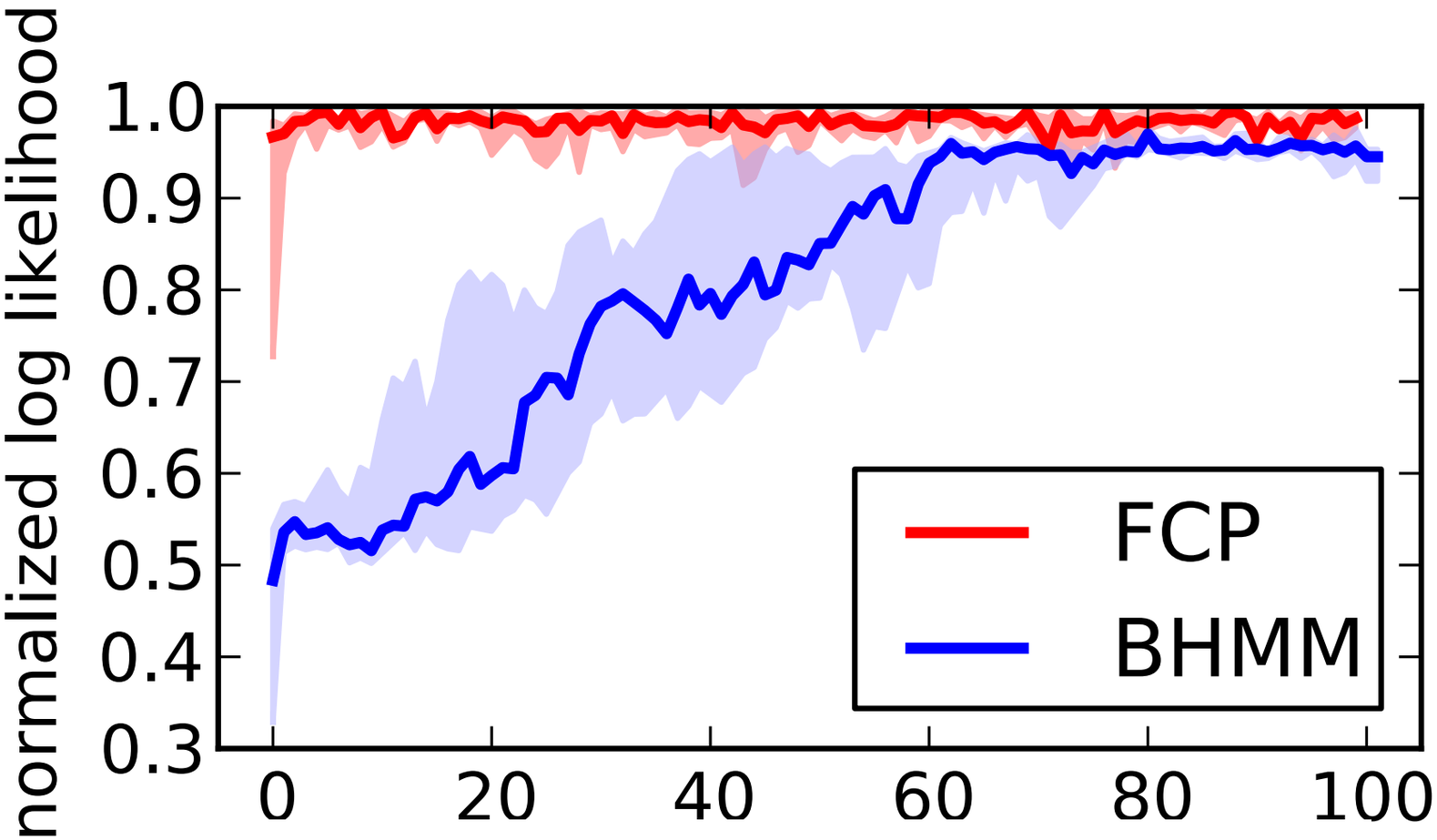
0000000000000000

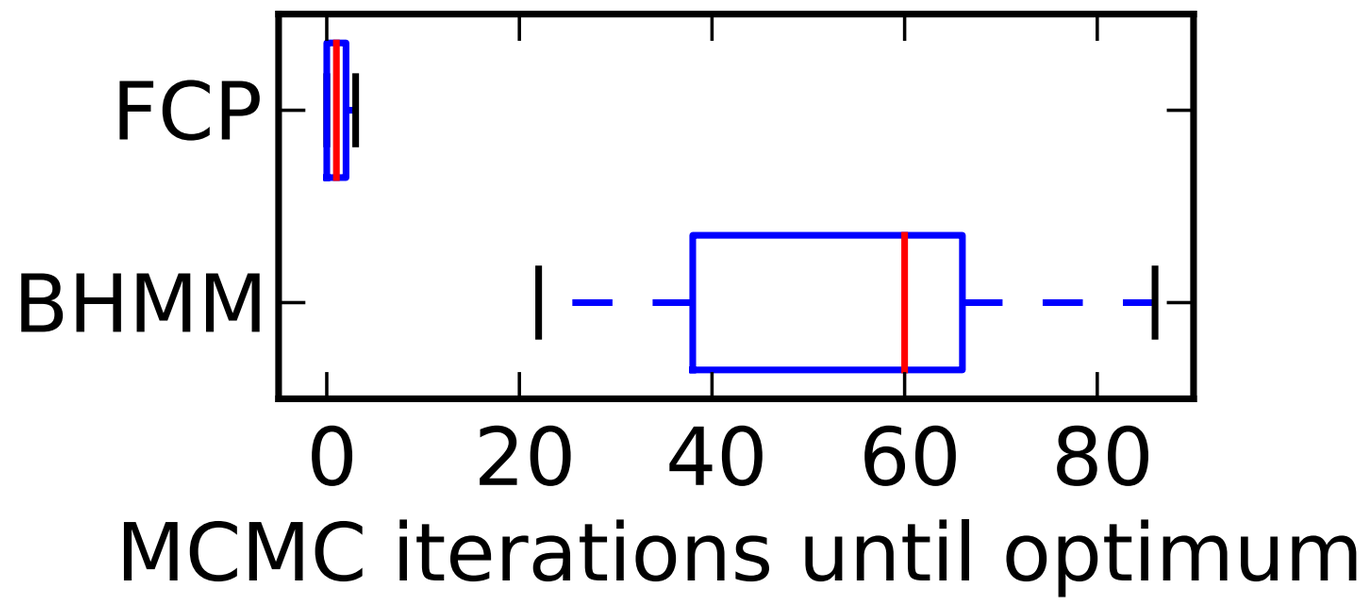
....

1111111111111111

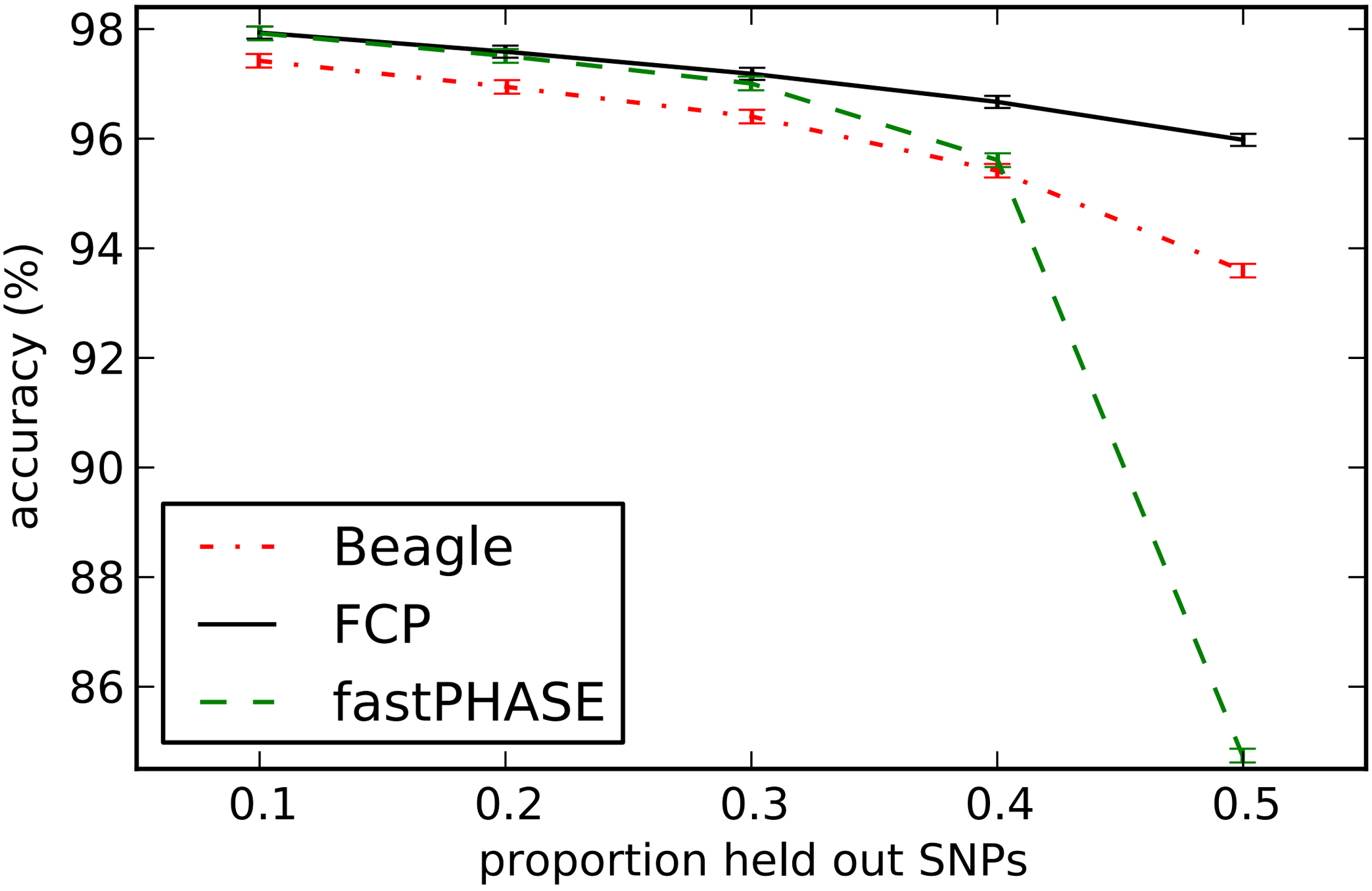
1111111111111111

....

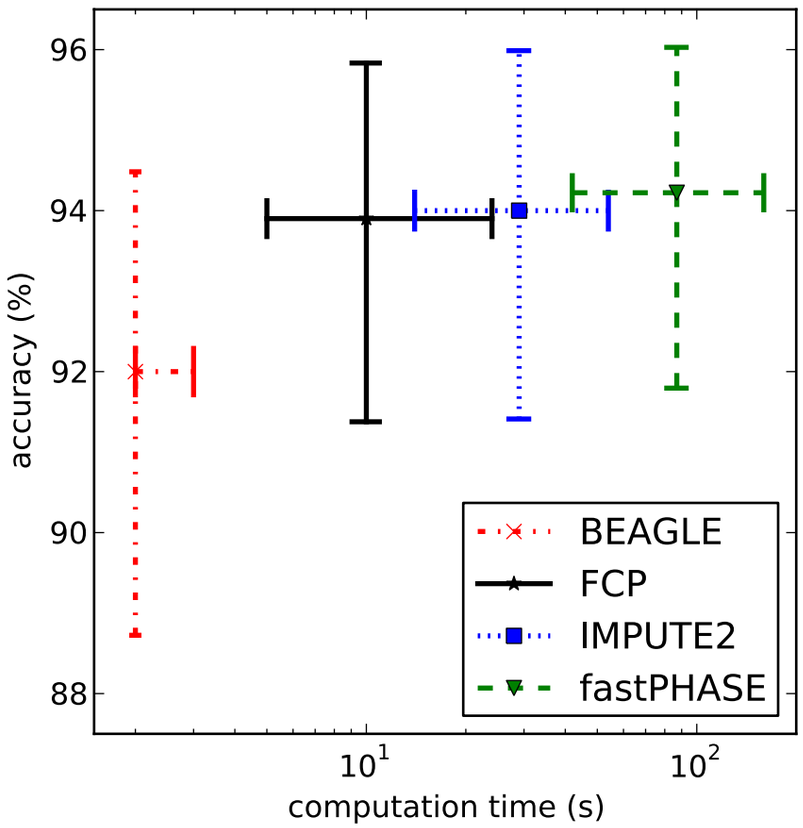


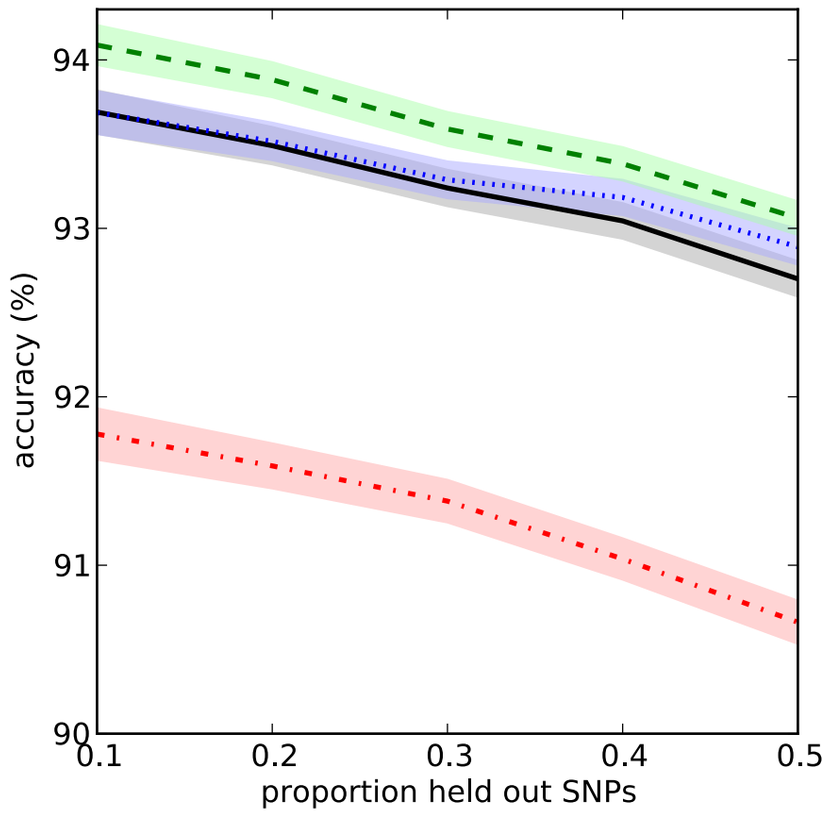


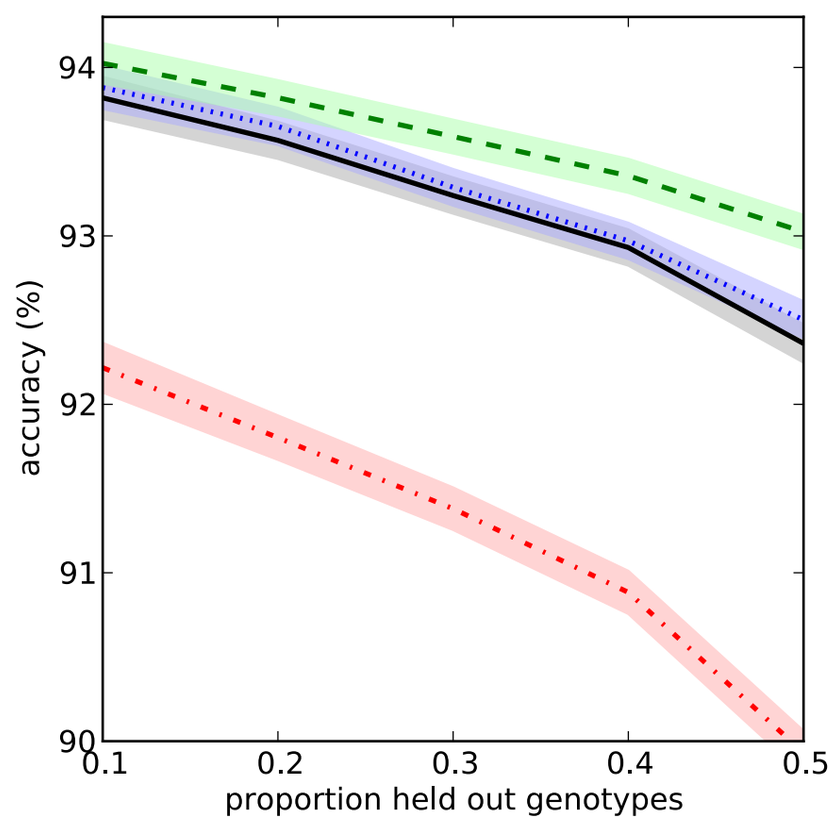
Genotype Imputation---Pre-phased data



Genotype Imputation---Unphased data







A Few Final Words

Summary

- Introduction to Bayesian learning and Bayesian nonparametrics.
- Dirichlet processes:
  - Infinite limit of finite mixture models.
  - Chinese restaurant processes, stick-breaking construction.
  - Ferguson's Definition
- Pitman-Yor processes:
  - Two-parameter Chinese restaurant processes.
  - Power-law properties.
- Hierarchical Bayesian nonparametric models.
- Infinite hidden Markov models.
- Random partitions, coagulations, fragmentations, trees.

What Were Not Covered Here

- Gaussian processes
- Indian buffet processes, beta processes.
- Other nonparametric dynamical models.
- Dependent random measures.
- Completely random measures and other generalizations of DPs.
- Combinatorial stochastic processes and their relationship to data structures and programming languages.
- Relational models, topic models etc.
- Foundational issues, convergence and asymptotics.

Future of Bayesian Nonparametrics

- Augmenting the standard modelling toolbox of machine learning.
- Development of better inference algorithms and software toolkits.
- Exploration of novel stochastic processes.
- More applications in machine learning and beyond.