Bayesian Nonparametrics: Random Partitions

Yee Whye Teh
Gatsby Computational Neuroscience Unit, UCL

Machine Learning Summer School Singapore
June 2011

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) 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: focus on the role of random partitions.

Bayesian Modelling

Probabilistic Machine Learning

- Machine Learning is all about data.
  - Stochastic process
  - Noisily observed
  - Partially observed
- **Probability theory** is a language to express all these notions.
  - **Probabilistic models**
- Allow us to reason coherently about such data.
- Give us powerful computational tools to implement such reasoning on computers.

Probabilistic Models

- 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.

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

- 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 Models

- Learning, typically by maximum likelihood:

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

- Prediction:

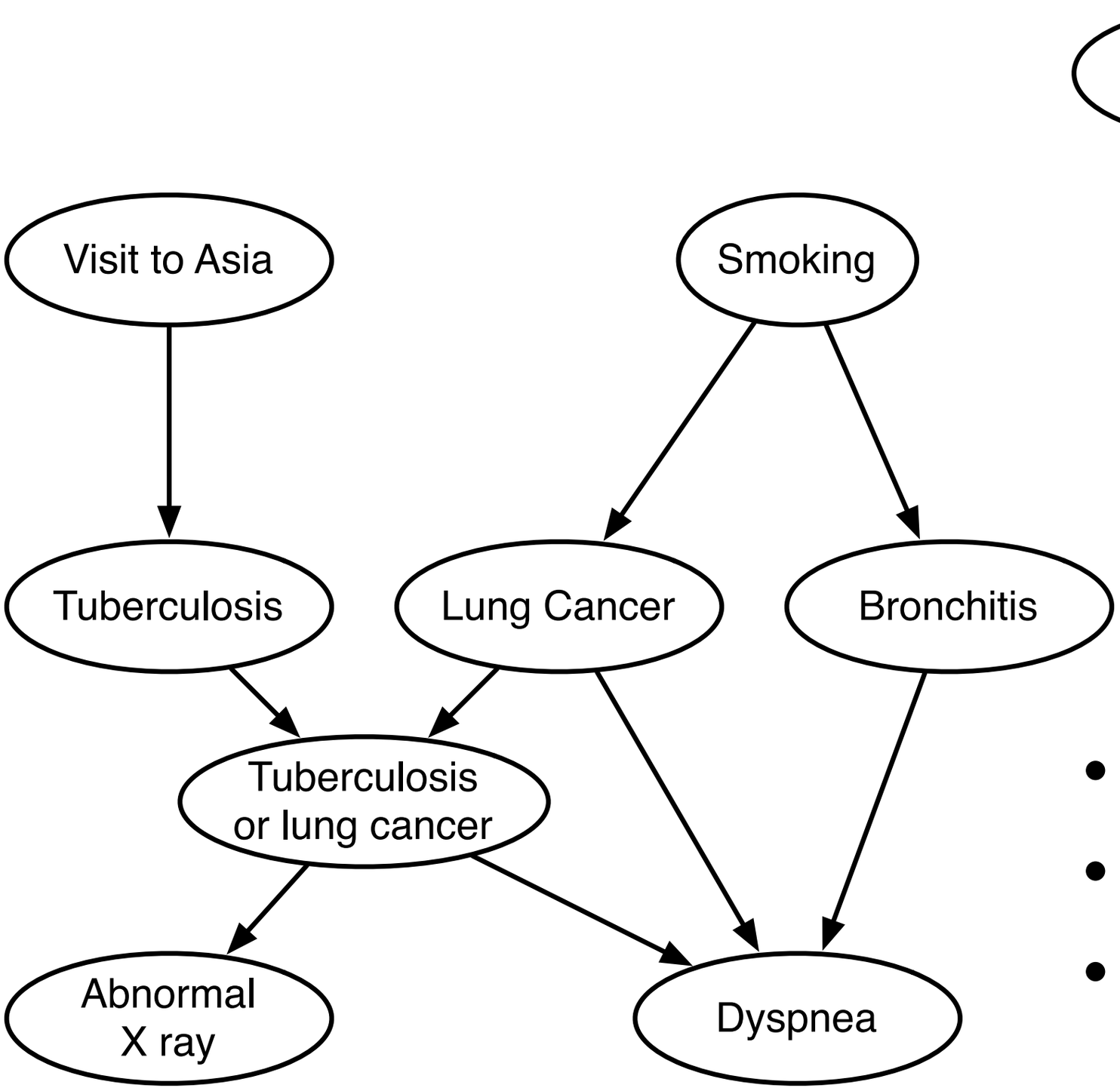
$$P(x_{n+1}, y_{n+1} | x_1, \ldots, x_n, \theta)$$

- Classification:

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

- Visualization, interpretation, summarization.

Graphical Models



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

Bayesian Machine Learning

- Prior distribution:

$$P(\theta)$$

- Posterior distribution (inference and learning):

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

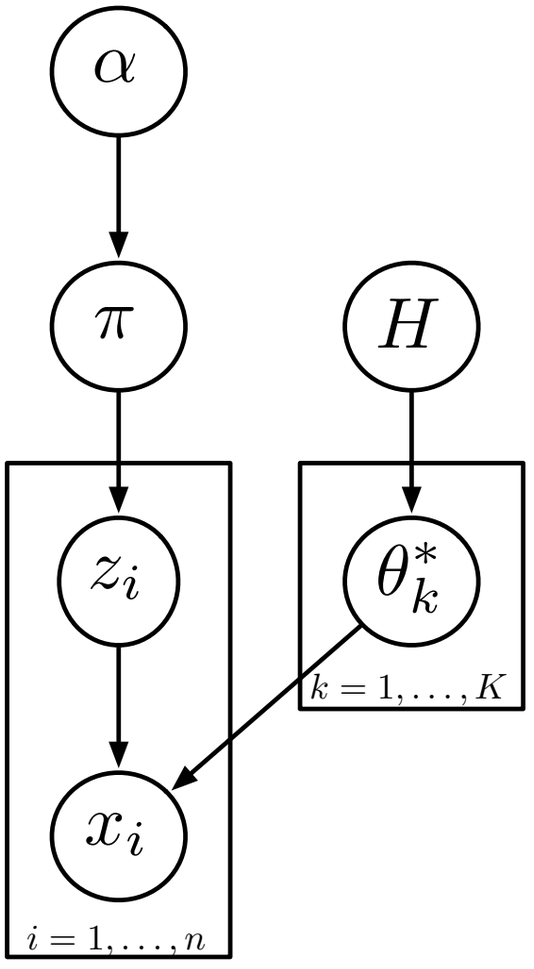
- Prediction:

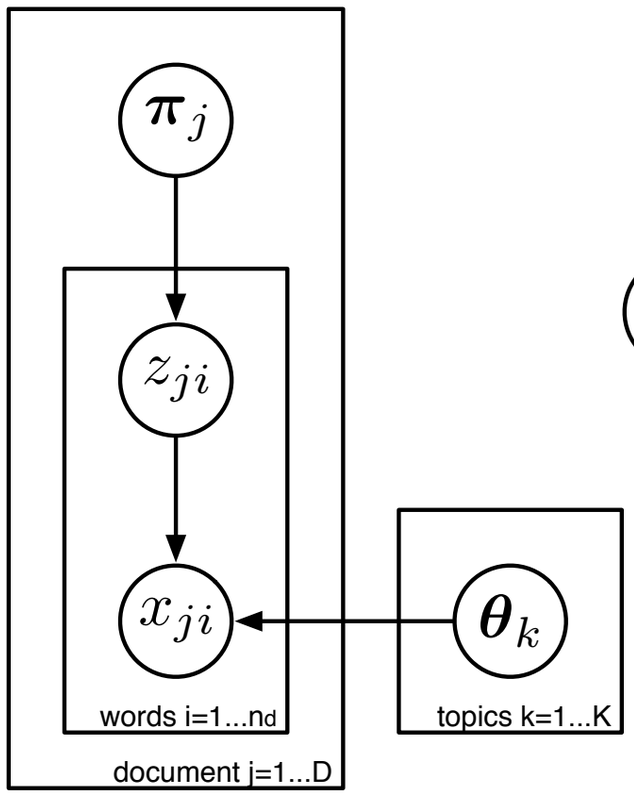
$$P(x_{n+1} | x_1, \ldots, x_n) = \int P(x_{n+1} | \theta) P(\theta | x_1, \ldots, x_n) d\theta$$

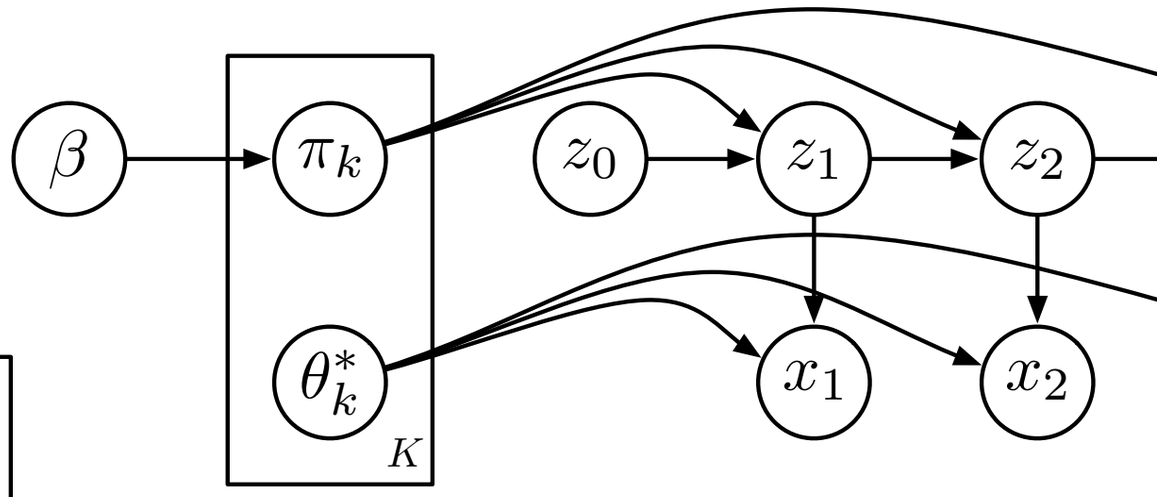
- Classification:

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

Bayesian Models

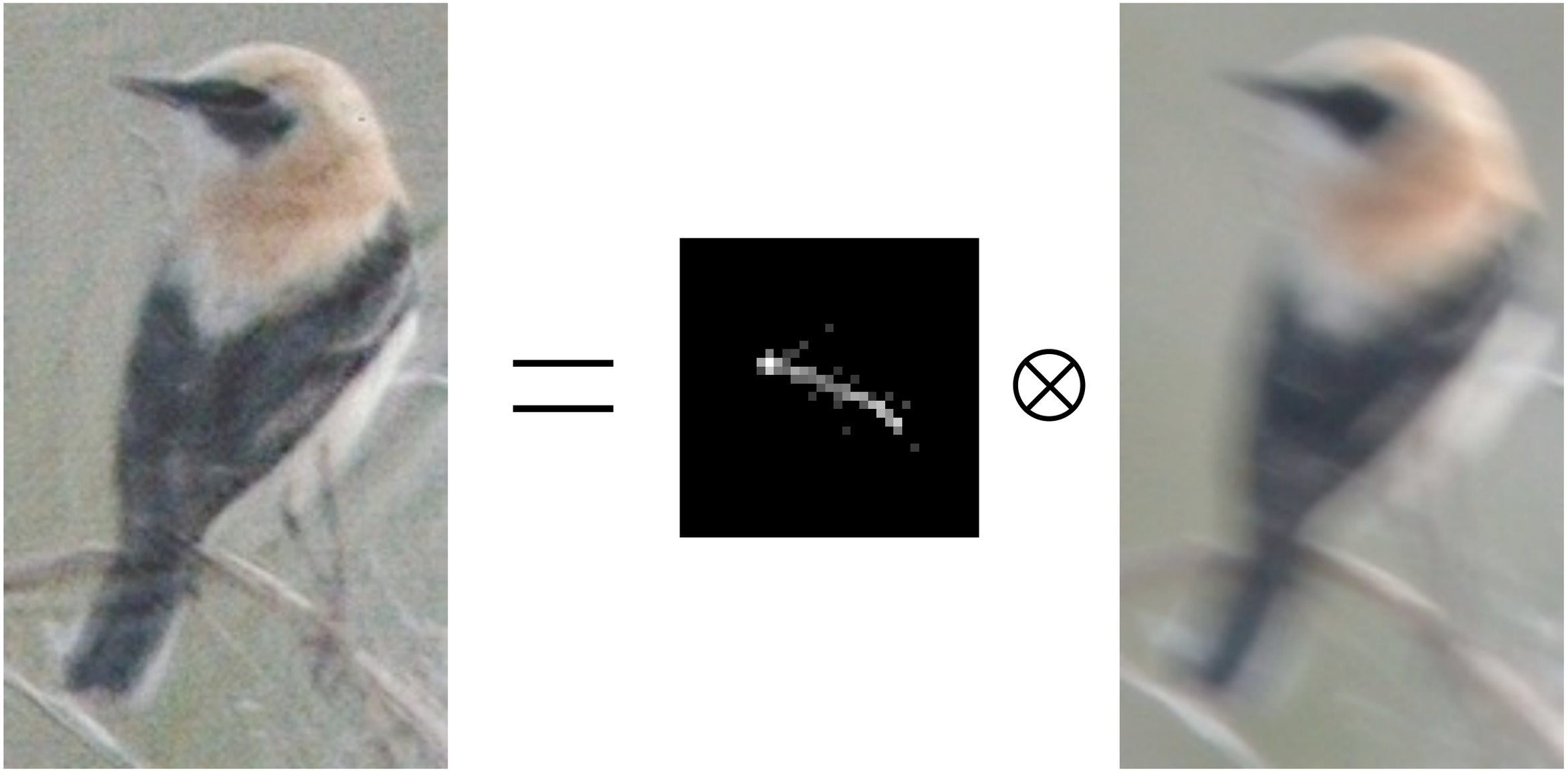






- Latent variables and parameters are not distinguished.
- Important operations are marginalization and posterior computation.

Blind Deconvolution



[Fergus et al 2006, Levin et al 2009]

Pros and Cons

- Maximum likelihood: have to guard against overfitting.
- Bayesian methods do not fit any parameters (no overfitting).
- Bayesian posterior distribution captures more information from data.
- Bayesian learning is coherent (no Dutch book).
- Prior distributions and model structures can be useful way to introduce prior knowledge.
- Bayesian inference is often more complex.
- Bayesian inference is often more computationally intensive.
- Powerful computational tools have been developed for Bayesian inference.
- Often easier to think about the “ideal” and “right way” to perform learning and inference before considering how to do it efficiently.

Bayesian Nonparametrics

[Hjort et al 2010]

Model Selection

- In non-Bayesian methods model selection is needed to prevent overfitting and underfitting.
- In Bayesian contexts model selection is also useful as a way of determining the appropriateness of various models given data.
- Marginal likelihood:

$$P(\mathbf{x}|M_k) = \int P(\mathbf{x}|\theta_k, M_k)P(\theta_k|M_k)d\theta_k$$

- Model selection:

$$\underset{M_k}{\operatorname{argmax}}\, P(\mathbf{x}|M_k)$$

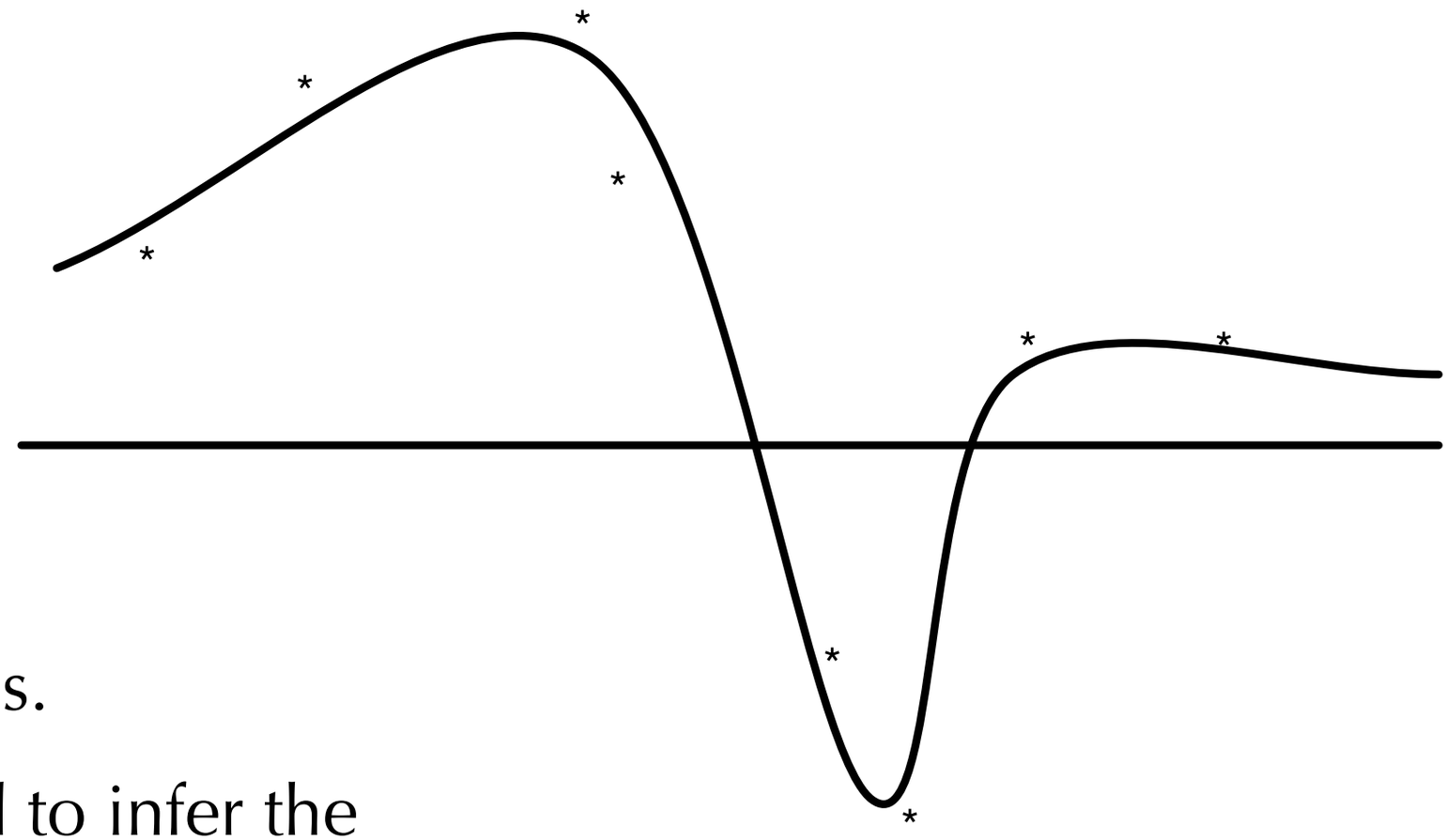
- Model averaging:

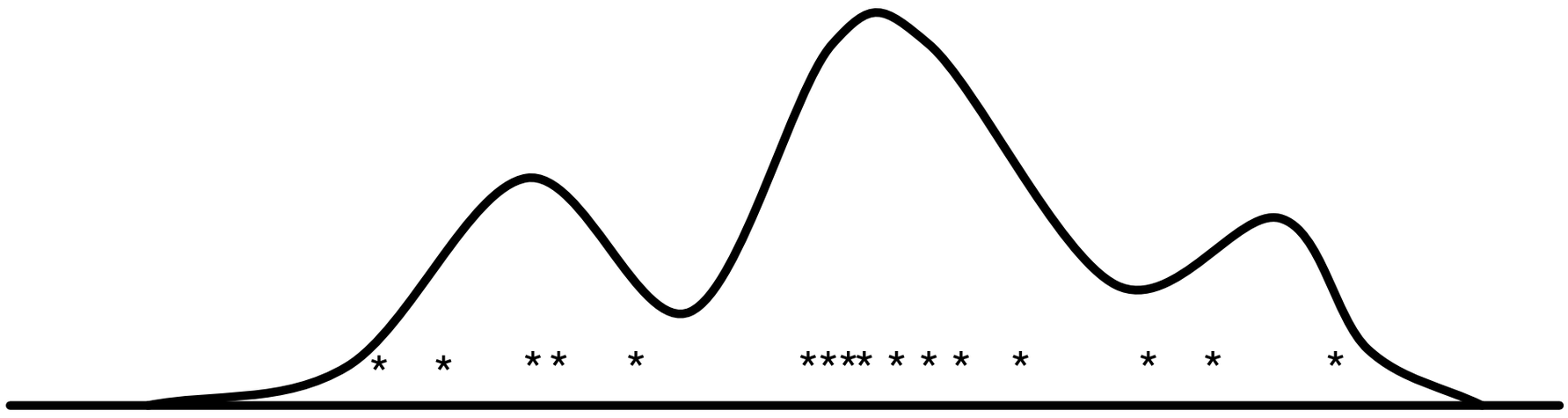
$$P(M_k|\mathbf{x}) \propto P(\mathbf{x}|M_k)P(M_k)$$

Model Selection

- Model selection is often very computationally intensive.
- But reasonable and proper Bayesian methods should not overfit anyway.
- Idea: use one large model, and be Bayesian so will not overfit.
- Bayesian nonparametric idea: use a very large Bayesian model avoids both overfitting and underfitting.

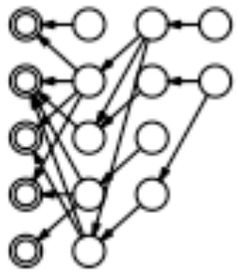
Large Function Spaces

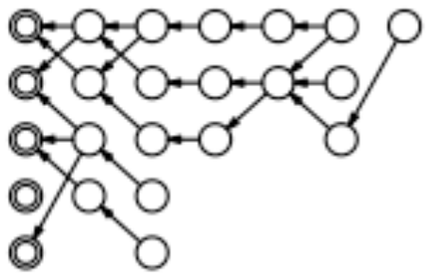


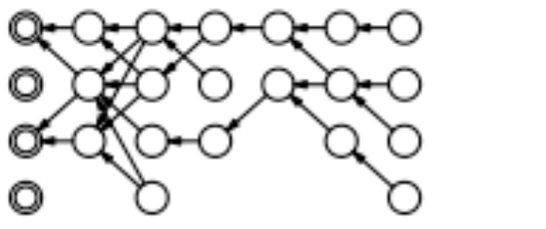


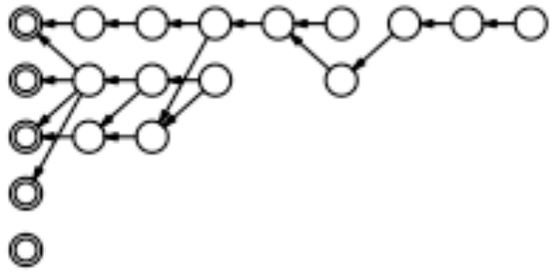
- Large function spaces.
- More straightforward to infer the infinite-dimensional objects themselves.

Structural Learning

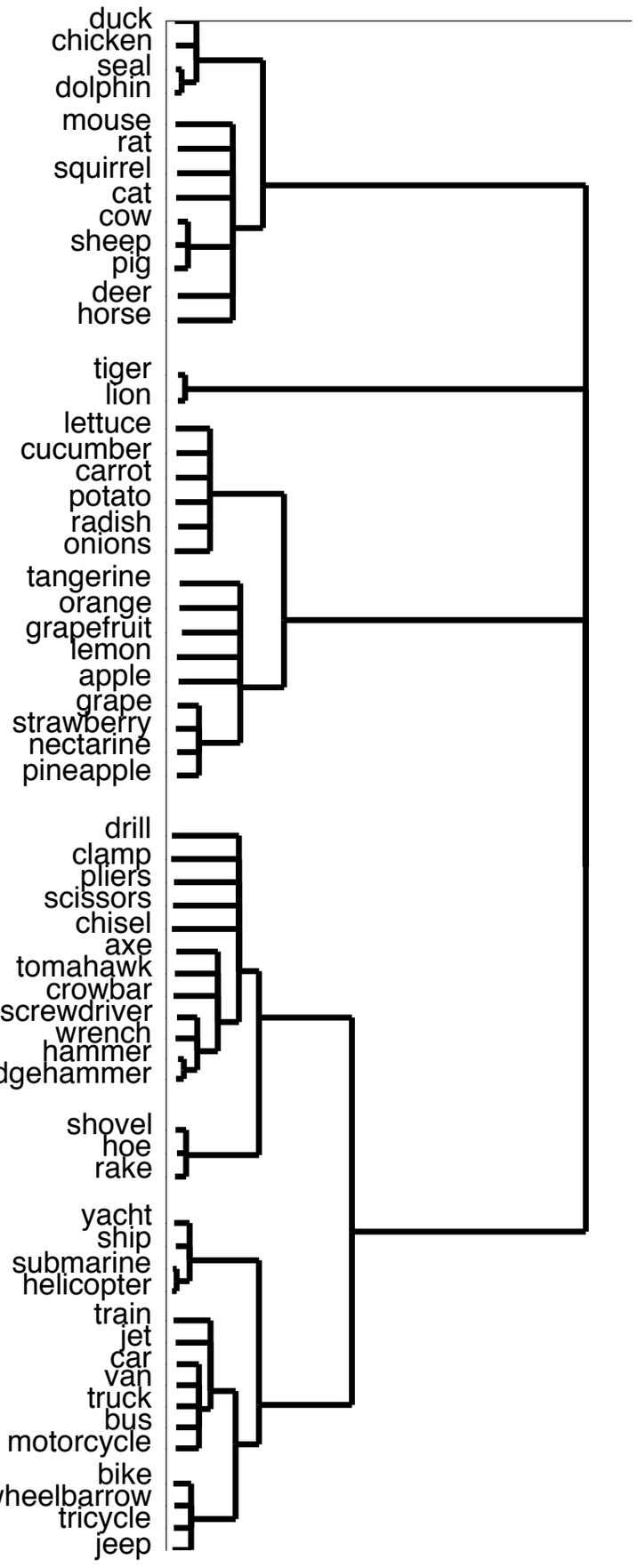








- Learning structures.
- Bayesian prior over combinatorial structures.
- Nonparametric priors sometimes end up simpler than parametric priors.



[Adams et al AISTATS 2010, Blundell et al UAI 2010]

Novel Models with 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).
- Techniques developed for Bayesian nonparametric models are applicable to many stochastic processes not traditionally considered as Bayesian nonparametric models.

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*.

Random Partitions in Bayesian Nonparametrics

Overview

- Random partitions, through their relationship to Dirichlet and Pitman-Yor processes, play very important roles in Bayesian nonparametrics.

- Introduce Chinese restaurant processes via finite mixture models.

- Projectivity, exchangeability, de Finetti's Theorem, and Kingman's paintbox construction.

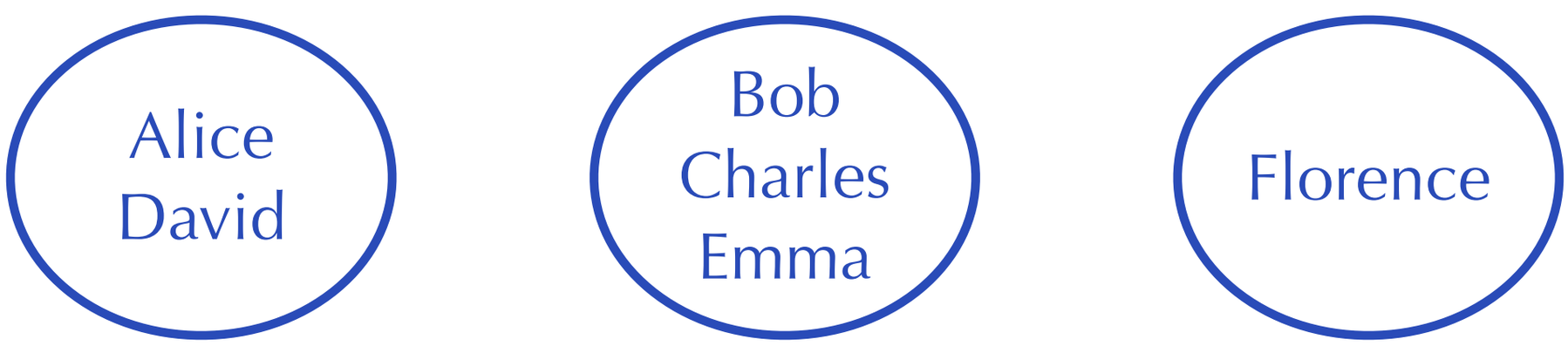
- Two parameter extension and the Pitman-Yor process.

- Infinite mixture models and inference in them.

Random Partitions

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 $\mathcal{P}_S$.

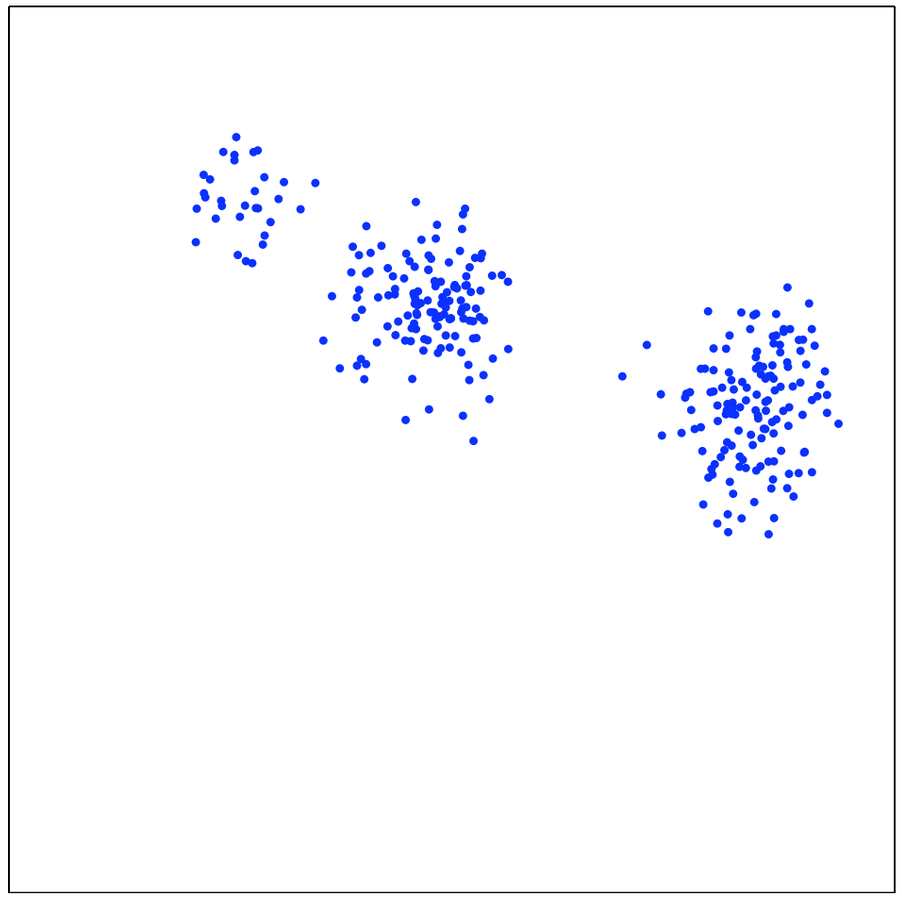
Partitions in Model-based Clustering

- Given a dataset S, partition it into clusters of similar items.
- **Cluster** $c \in \varrho$ described by a model

$$F(x|\theta_c)$$

parameterized by θ_c.

- Bayesian approach: introduce prior over ϱ and θ_c; compute posterior over both.
- To do so we will need to work with **random partitions**.

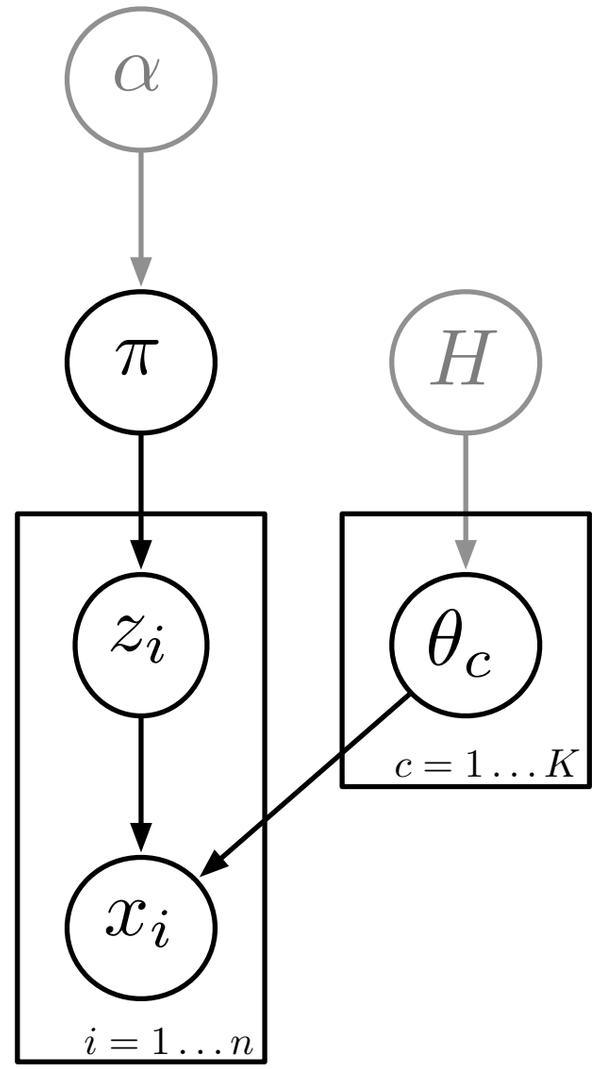


From Finite Mixture Models to Chinese Restaurant Processes

Finite Mixture Models

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

$$\pi|\alpha \sim \text{Dirichlet}(\alpha)$$
$$\theta_c|H \sim H$$



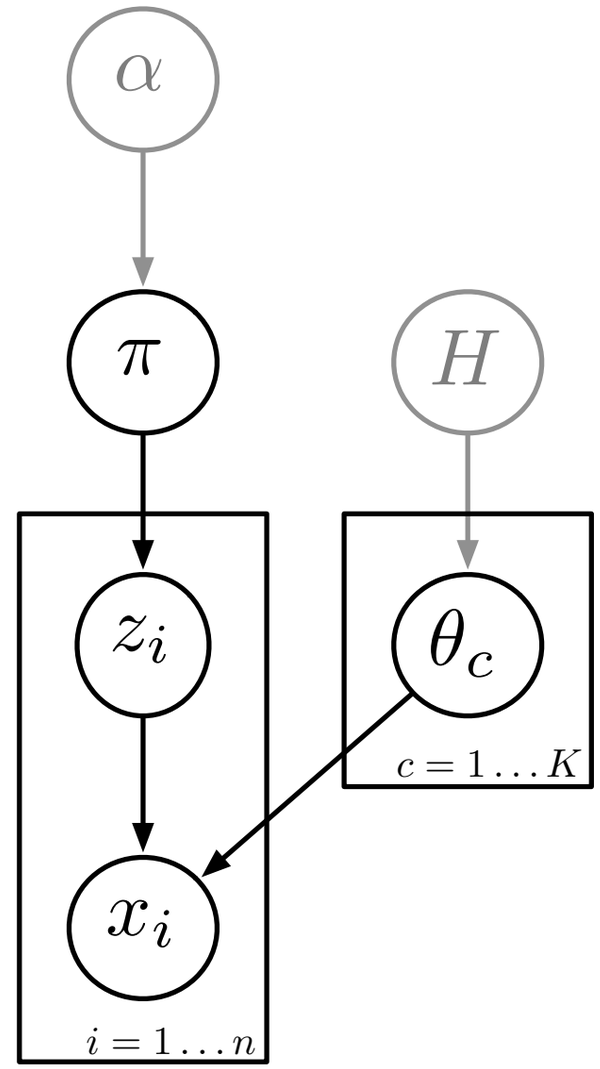
Finite Mixture Models

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

$$P(\pi|\alpha) = \frac{\Gamma(\sum_c \alpha_c)}{\prod_c \Gamma(\alpha_c)} \prod_{c=1}^{K} \pi_c^{\alpha_c - 1}$$

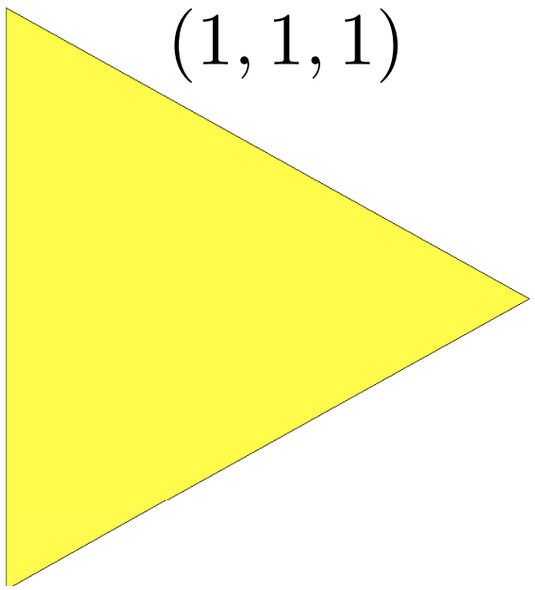
  with $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^x dx$ and $\alpha_c \geq 0$.

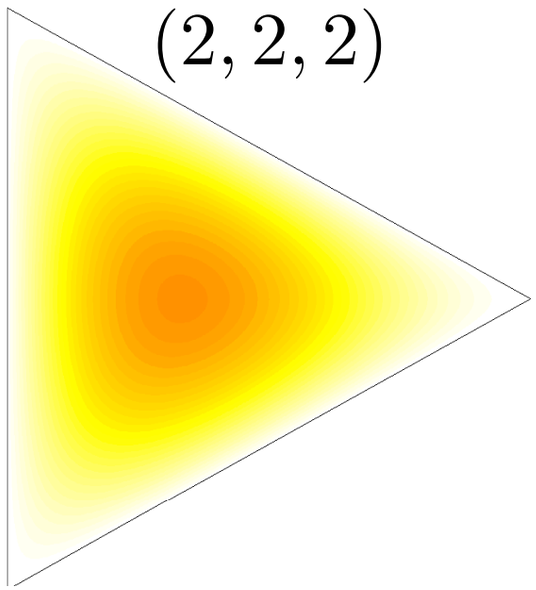
- Standard distribution on probability vectors, due to **conjugacy** with multinomial.

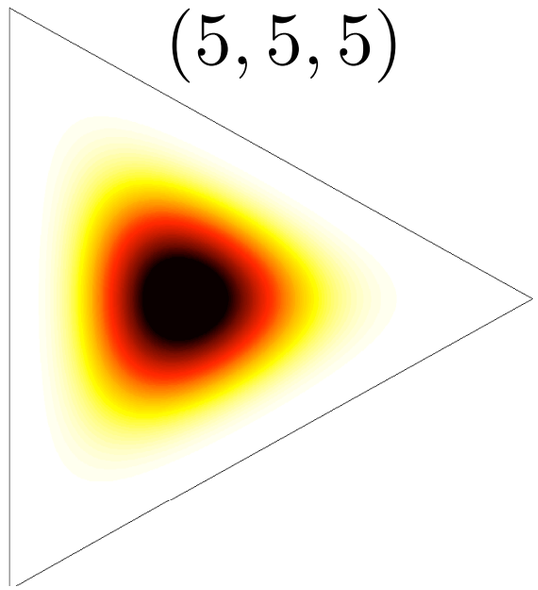


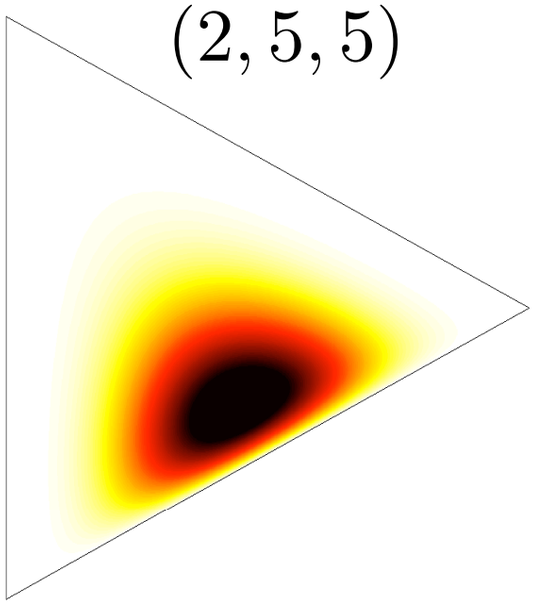
$$\pi|\alpha \sim \text{Dirichlet}(\alpha)$$
$$\theta_c|H \sim H$$
$$z_i|\pi \sim \text{Multinomial}(\pi)$$
$$x_i|z_i, \theta_{z_i} \sim F(\cdot|\theta_{z_i})$$

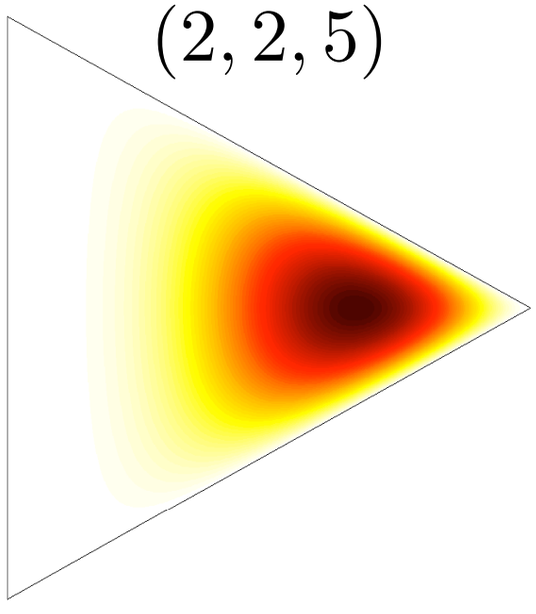
Dirichlet Distribution

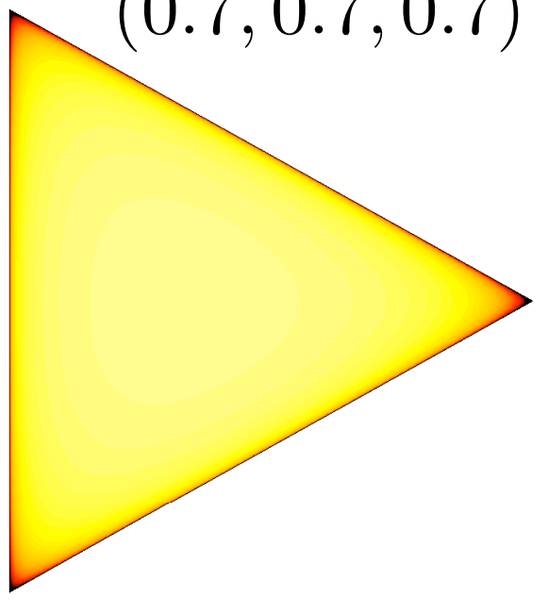












$$P(\pi|\alpha) = \frac{\Gamma(\sum_c \alpha_c)}{\prod_c \Gamma(\alpha_c)} \prod_{c=1}^{K} \pi_c^{\alpha_c - 1}$$

Dirichlet-Multinomial Conjugacy

- Joint distribution over z_i and π:

$$P(\pi|\alpha) \times \prod_{i=1}^{n} P(z_i|\pi) = \frac{\Gamma(\sum_c \alpha_c)}{\prod_c \Gamma(\alpha_c)} \prod_{c=1}^{K} \pi_c^{\alpha_c - 1} \times \prod_{c=1}^{K} \pi_c^{n_c}$$

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

- Posterior distribution:

$$P(\pi|\mathbf{z}, \alpha) = \frac{\Gamma(n + \sum_c \alpha_c)}{\prod_c \Gamma(n_c + \alpha_c)} \prod_{c=1}^{K} \pi_c^{n_c + \alpha_c - 1}$$

- Marginal distribution:

$$P(\mathbf{z}|\alpha) = \frac{\Gamma(\sum_c \alpha_c)}{\prod_c \Gamma(\alpha_c)} \frac{\prod_c \Gamma(n_c + \alpha_c)}{\Gamma(n + \sum_c \alpha_c)}$$

Induced Distribution over Partitions

$$P(\mathbf{z}|\alpha) = \frac{\Gamma(\sum_c \alpha_c)}{\prod_c \Gamma(\alpha_c)} \frac{\prod_c \Gamma(n_c + \alpha_c)}{\Gamma(n + \sum_c \alpha_c)}$$

- 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 of the data set (without labelling).
- Start by supposing $\alpha_c = \alpha/K$.
- The partition ϱ has $|\varrho| = k \leq K$ clusters, each of which can be assigned one of K labels (without replacement). So after some algebra:

$$P(\mathbf{z}|\alpha) = \frac{1}{K(K-1)\cdots(K-k+1)} P(\varrho|\alpha)$$

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

Chinese Restaurant Process

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

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

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

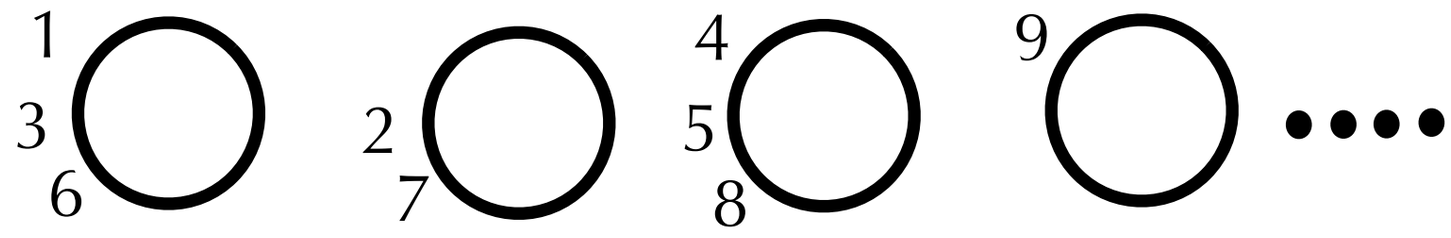
(using $\Gamma(\alpha/K)\alpha/K = \Gamma(1+\alpha/K)$).

- This is the **Chinese restaurant process** (CRP).
- We write $\varrho \sim \mathrm{CRP}([n], \alpha)$ if $\varrho \in \mathcal{P}_{[n]}$ is CRP distributed.

Chinese Restaurant Processes

[Aldous 1985, Pitman 2006]

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 all terms together, we get the overall probability of ϱ:

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

Exchangeability

Exchangeability

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

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

Exchangeability

- 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 $S = [9] = \{1,\ldots,9\}$, and σ is a permutation of $[9]$.

Exchangeability

- 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 $S = [9] = \{1,\ldots,9\}$, and σ is a permutation of $[9]$.

- The Chinese restaurant process satisfies exchangeability:

Exchangeability

- 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).

Exchangeability

- 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.

Exchangeability

- 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.

Consistency and Projectivity

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:

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

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:

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

- A sequence of distributions $P_1, P_2, \ldots$ over $\mathcal{P}_{[1]}, \mathcal{P}_{[2]}, \ldots$ is **projective** or **consistent** if the distribution on $\mathcal{P}_{[n]}$ induced by P_m for $m>n$ is P_n.

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:

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

- A sequence of distributions $P_1, P_2, \ldots$ over $\mathcal{P}_{[1]}, \mathcal{P}_{[2]}, \ldots$ is **projective** or **consistent** if the distribution on $\mathcal{P}_{[n]}$ induced by P_m for $m>n$ is P_n.

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

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:

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

- A sequence of distributions $P_1, P_2, \ldots$ over $\mathcal{P}_{[1]}, \mathcal{P}_{[2]}, \ldots$ is **projective** or **consistent** if the distribution on $\mathcal{P}_{[n]}$ induced by P_m for $m>n$ is P_n.

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

- The Chinese restaurant process is projective since:

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:

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

- A sequence of distributions $P_1, P_2, \ldots$ over $\mathcal{P}_{[1]}, \mathcal{P}_{[2]}, \ldots$ is **projective** or **consistent** if the distribution on $\mathcal{P}_{[n]}$ induced by P_m for $m>n$ is P_n.

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

- The Chinese restaurant process is projective since:
  - The finite mixture model is, and also it is defined sequentially.

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:

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

- A sequence of distributions $P_1, P_2, \ldots$ over $\mathcal{P}_{[1]}, \mathcal{P}_{[2]}, \ldots$ is **projective** or **consistent** if the distribution on $\mathcal{P}_{[n]}$ induced by P_m for $m>n$ is P_n.

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

- 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).

Projective and Exchangeable Partitions

- Projective and exchangeable random partitions over [n] can be extended to distributions over partitions of $\mathbb{N}$ in a unique manner.
- *Can we characterize all random exchangeable projective partitions of* $\mathbb{N}$*?*

Projective and Exchangeable Partitions

- Let G be a distribution, with possibly both atomic and smooth components:

$$G = G_0 + \sum_{c=1}^{\infty} \pi_c \delta_{\theta_c}$$

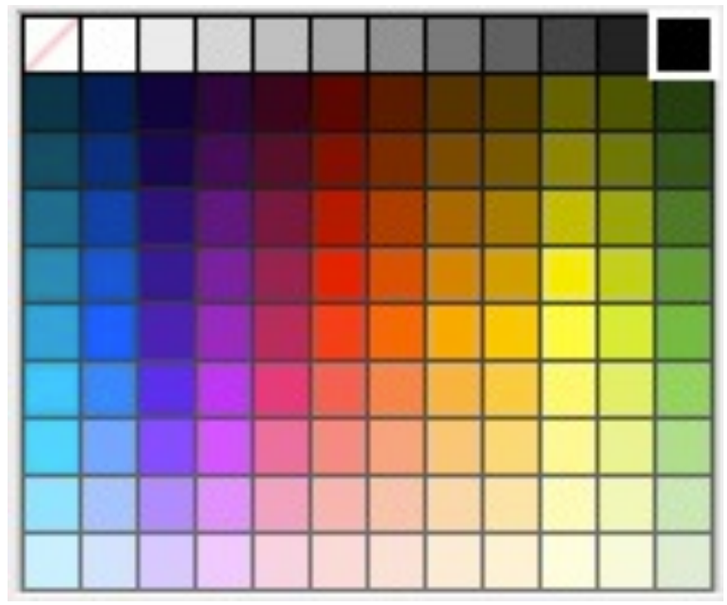
- Let ϕi ~ G for each i independently, and define an exchangeable random partition ϱ based on the unique values of {ϕi}.
- If G is random, the distribution over ϱ is given by

$$P(\varrho) = \int P(\varrho|G)P(G)dG$$

- *All exchangeable random partitions admit such a representation.*
- Note that the integral is written as if both ϱ and G have densities; this is not (typically) true.

Kingman's Paint-box Construction

- A paint-box consists of a number of colours, with colour c picked with probability π_c. There is also a special colour 0 with probability π_0 which looks different each time it is picked.
  - For each $i \in \mathbb{N}$ pick a colour from the paint-box.
  - This defines a partition ϱ of $\mathbb{N}$ according the different colours picked.
- Given a distribution over the probabilities $\{\pi_c\}$ we get an exchangeable random ϱ.



[Kingman JRSSB 1975]

De Finetti's Theorem

- Let $x_1, x_2, x_3, \ldots$ be an **infinitely exchangeable** (i.e. projective and 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].

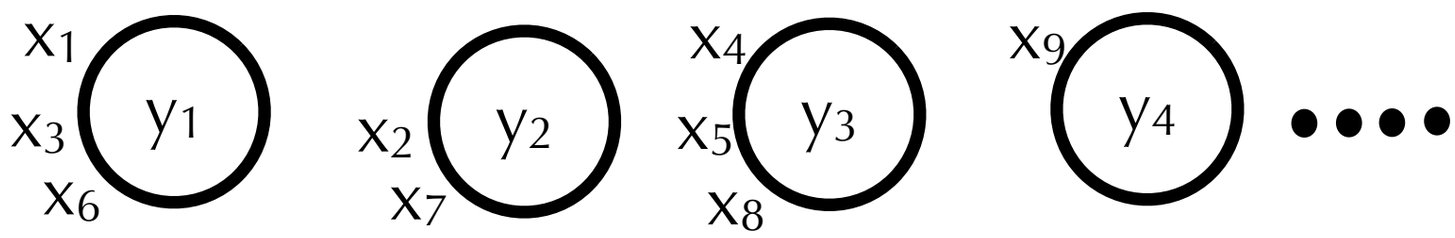
- Then there is a latent variable G such that:

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

[De Finetti 1931, Kallenberg 2005]

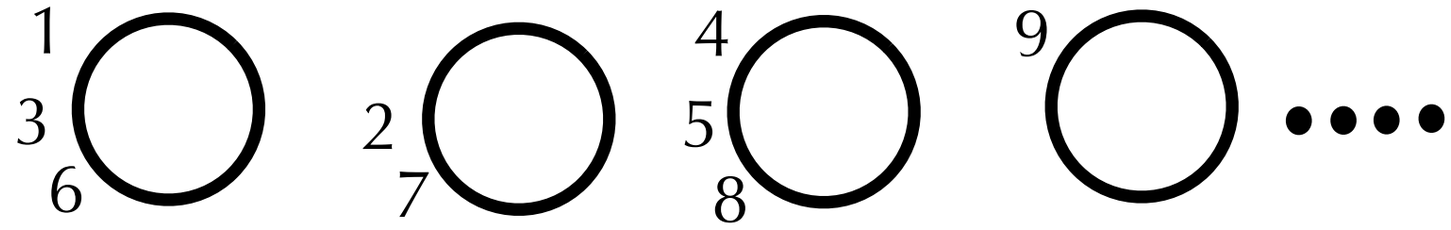
Dirichlet Process

- Since the Chinese restaurant process is exchangeable, we can define an an infinitely exchangeable sequence as follows:
  - Sample $\varrho \sim \text{CRP}(\mathbb{N}, \alpha)$.
  - For $c \in \varrho$:
    - sample $y_c \sim H$.
  - For i=1,2,...:
    - set $x_i = y_c$ where $i \in c$.
- The resulting de Finetti measure is the **Dirichlet Process** with parameters α and H (DP(α,H)).



[Ferguson AoS 1973, Blackwell & MacQueen AoS 1973]

Two-parameter Chinese Restaurant Processes



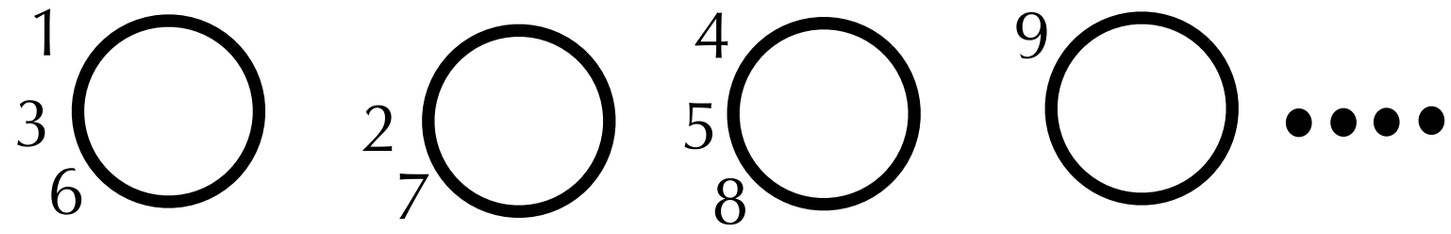
$$P(\text{sit at table } c) = \frac{n_c - d}{\alpha + \sum_{c \in \varrho} n_c} \qquad P(\text{sit at new table}) = \frac{\alpha + d|\varrho|}{\alpha + \sum_{c \in \varrho} n_c}$$

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

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

[Perman et al 1992, Pitman & Yor AoP 1997, Goldwater et al NIPS 2006, Teh ACL 2006]

Two-parameter Chinese Restaurant Processes



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

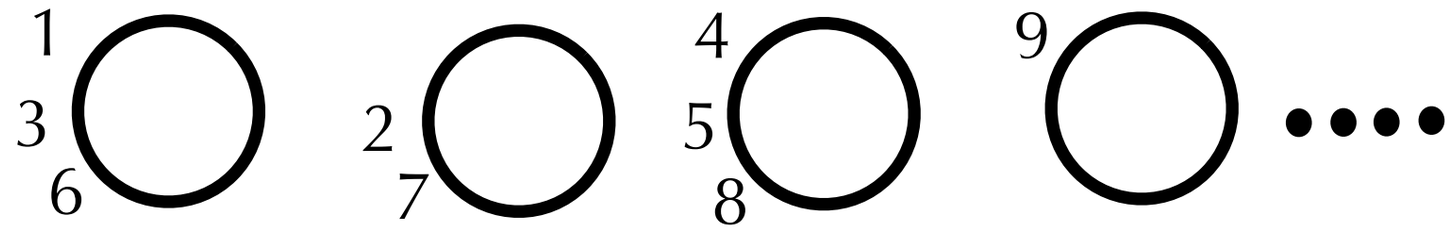
- The **two-parameter Chinese restaurant process** CRP([n],d,α) is a distribution over $\mathcal{P}_{[n]}$: ($0 \leq d < 1$, $\alpha > -d$)

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

- These are also projective and exchangeable distributions.

[Perman et al 1992, Pitman & Yor AoP 1997, Goldwater et al NIPS 2006, Teh ACL 2006]

Two-parameter Chinese Restaurant Processes



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

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

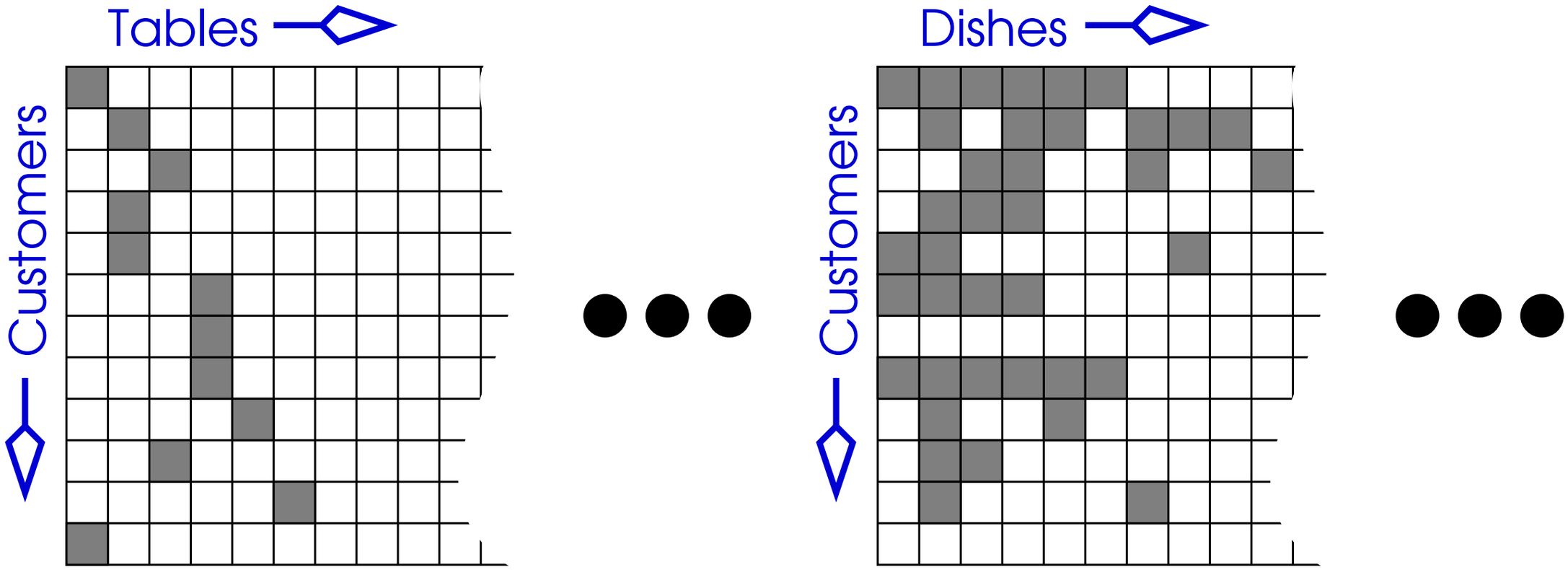
$$P(\varrho) = \frac{[\alpha + d]_d^{|\varrho|-1}}{[\alpha + 1]_1^{n-1}} \prod_{c \in \varrho} [1 - d]_1^{|c|-1} \qquad [z]_b^m = z(z + b) \cdots (z + (m - 1)b)$$

- These are also projective and exchangeable distributions.
- De Finetti measure is the **Pitman-Yor process**, which is a generalization of the Dirichlet process.

[Perman et al 1992, Pitman & Yor AoP 1997, Goldwater et al NIPS 2006, Teh ACL 2006]

Indian Buffet Processes

- Mixture models fundamentally use very simple representations of data:
  - Each data item belongs to just one cluster.
- Better representation if we allow multiple clusters per data item.
  - **Indian buffet processes** (IBPs).



[Griffiths & Ghahramani 2006]

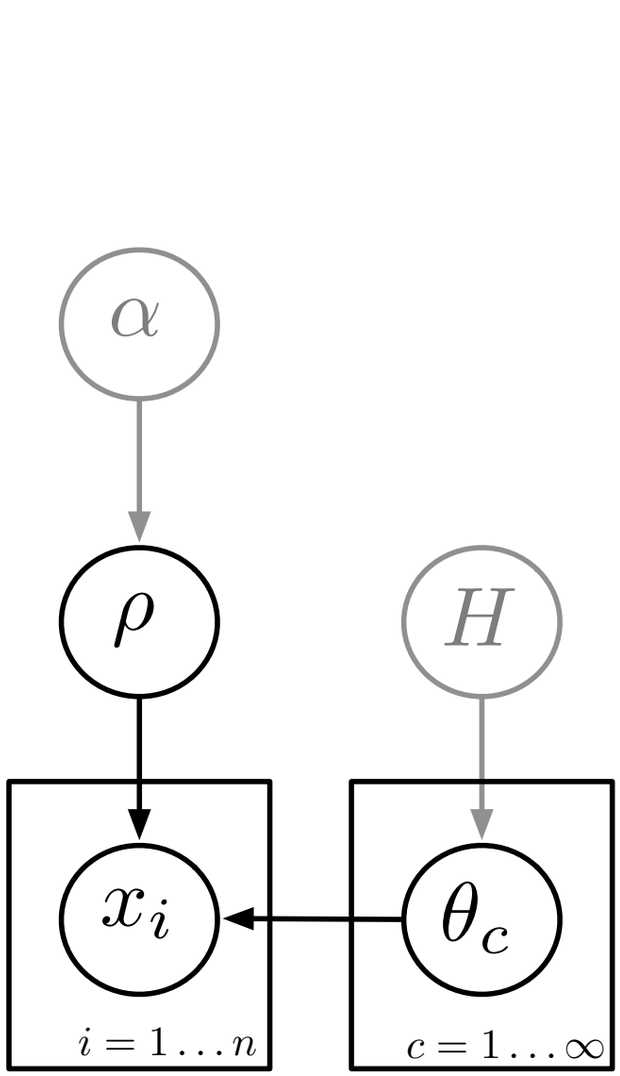
Infinite Mixture Models

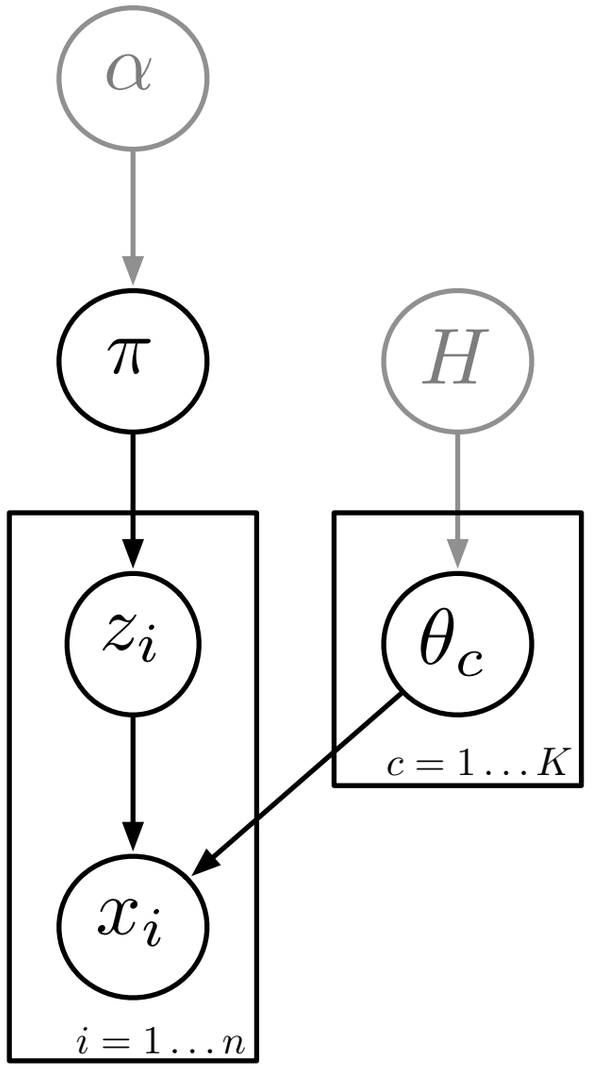
[Neal JCGS 2000, Rasmussen NIPS 2000, Ishwaran & Zarepour 2002]

Infinite Mixture Models

- Derived CRPs from finite mixture models.
- Taking K $\rightarrow \infty$ gives **infinite mixture models**.
- Expressed using CRPs:

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





Number of Clusters

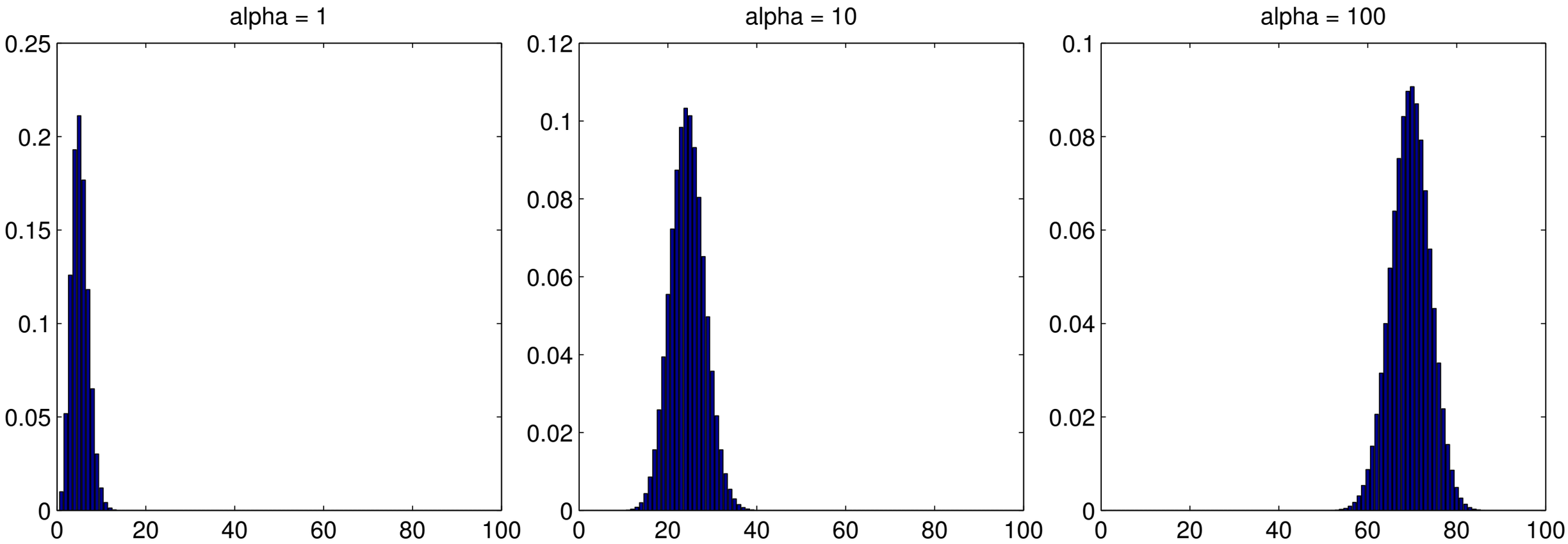
- Only a small number of mixture components used to model data.
- Prior over the number of clusters important in understanding the effect of the CRP prior (and through the infinite limit, the Dirichlet prior).
- Prior expectation and variance are:

$$E[k|n] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} = \alpha(\psi(\alpha + n) - \psi(\alpha)) \in O\left(\alpha \log\left(1 + \frac{n}{\alpha}\right)\right)$$

$$V[k|n] = \alpha(\psi(\alpha + n) - \psi(\alpha)) + \alpha^2(\psi'(\alpha + n) - \psi'(\alpha)) \in O\left(\alpha \log\left(1 + \frac{n}{\alpha}\right)\right)$$

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

Number of Clusters



- Prior number of clusters strongly dependent on α, and has small variance (not very uninformative).
- If uncertain about the number of clusters, need to place prior on α and marginalize over it.

Gibbs Sampling

- Iteratively resample clustering of each data item.

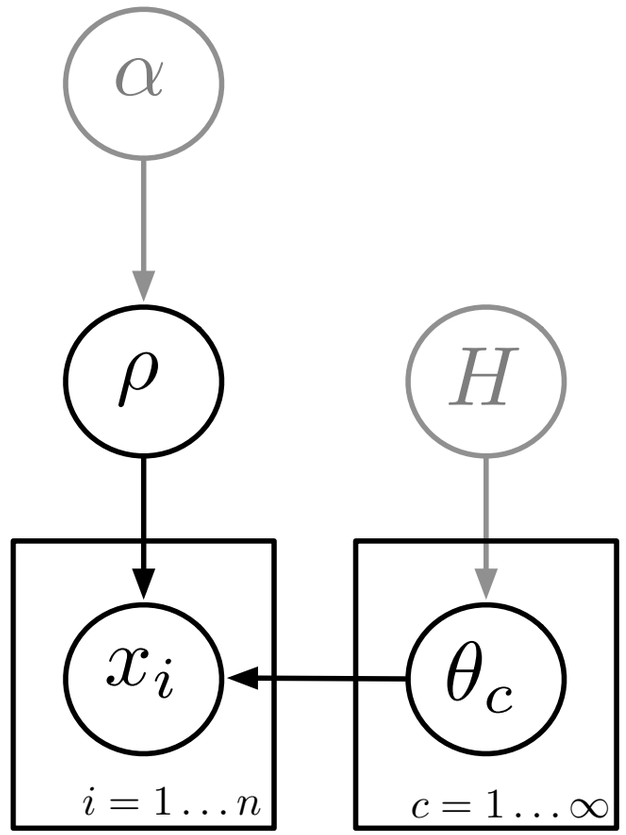
$$\rho_i = \text{cluster of } i$$

$$P(\rho) = \prod_{i=1}^{n} P(\rho_i | \rho_{1:i-1})$$

- Probability of cluster of data item i is:

$$P(\rho_i | \rho_{|\setminus i}, \mathbf{x}, \theta) = P(\rho_i | \rho_{|1:i-1}) \prod_{j=i+1}^{n} P(\rho_j | \rho_{|1:j-1}) \times P(x_i | \rho_i, \mathbf{x}_{\setminus i}, \theta_{\rho_i})$$

- Complex and expensive to compute.



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

Gibbs Sampling

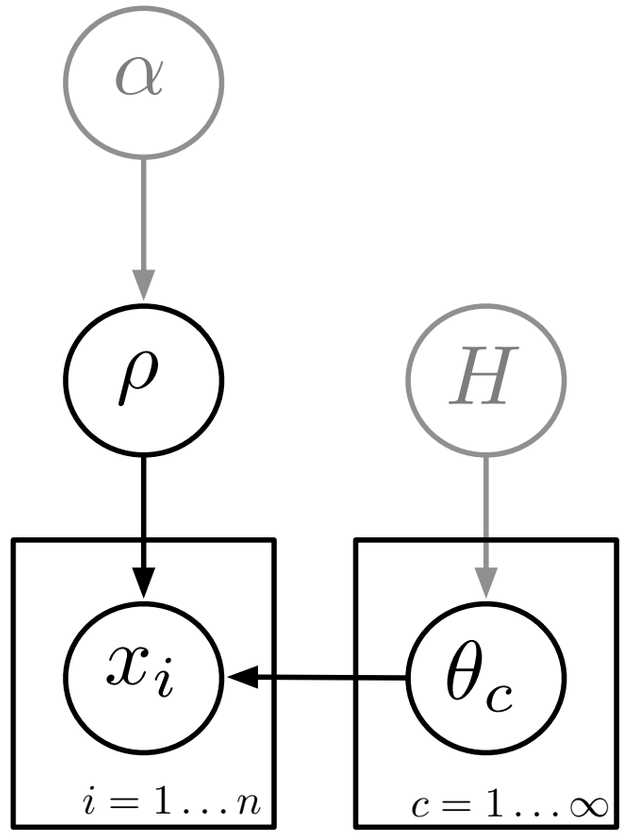
- Iteratively resample clustering of each data item.

- Make use of exchangeability of ϱ to treat data item i as the *last customer* entering restaurant.

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

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

- Simpler conditionals.



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

Parameter Updates

- Conditional distribution for α is:

$$P(\alpha|\rho) \propto P(\alpha)P(\rho|\alpha)$$

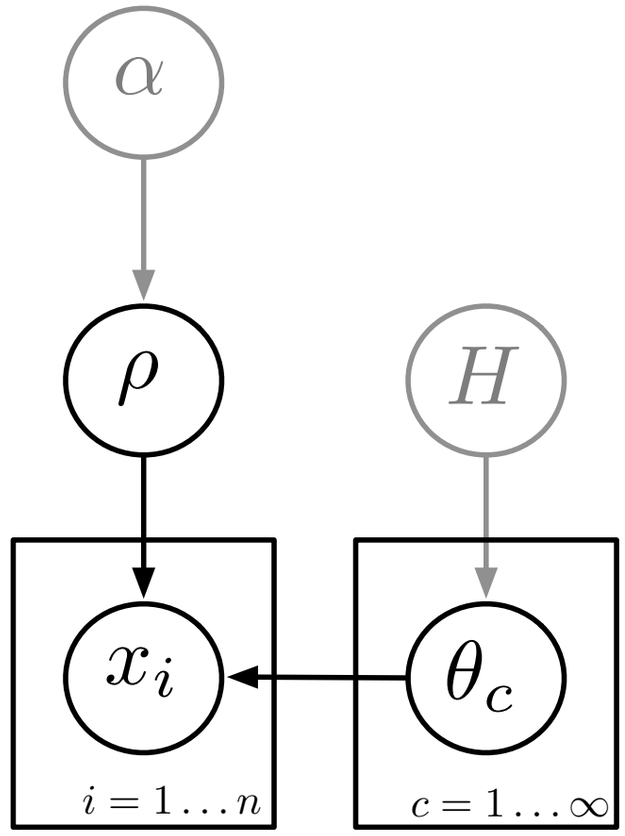
$$= P(\alpha)\frac{\alpha^{|\varrho|}\Gamma(\alpha)}{\Gamma(n+\alpha)}\prod_{c\in\varrho}\Gamma(|c|)$$

- Various updates: auxiliary variables Gibbs, Metropolis-Hastings, slice sampling.

- Typical prior for α is gamma distribution.

$$P(\alpha) \propto \alpha^{a-1}e^{-b\alpha}$$

$$\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} = \frac{1}{\Gamma(n)}\int_0^1 t^{\alpha-1}(1-t)^{n-1}dt$$



$$\rho|\alpha \sim \mathrm{CRP}([n], \alpha)$$

$$\theta_c|H \sim H \text{ for } c \in \rho$$

$$x_i|\theta_c \sim F(\cdot|\theta_c) \text{ for } c \ni i$$

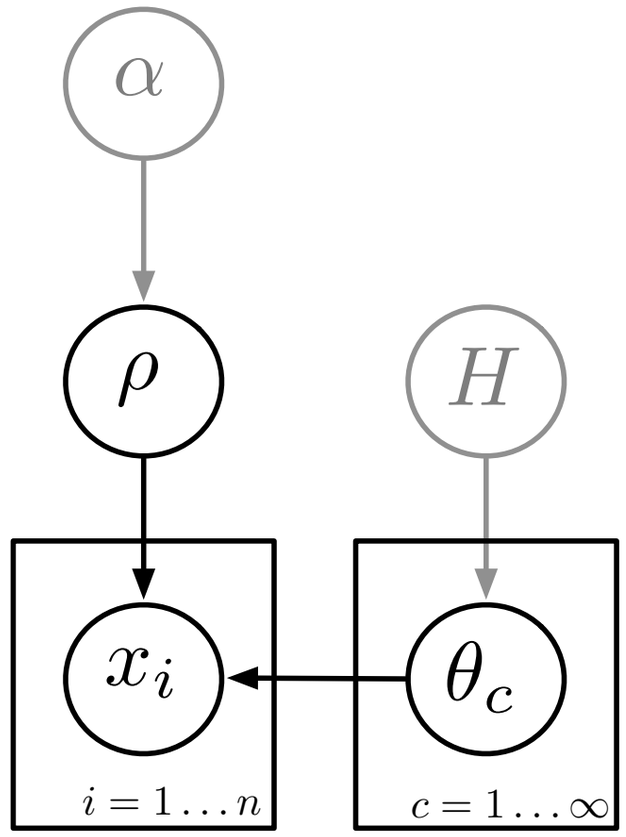
Parameter Updates

- Conditional distribution for θ_c are:

$$P(\theta_c) \propto H(\theta_c) \prod_{i \in c} F(x_i|\theta_c)$$

- If H is conjugate to P(x|θ), θ_c's can be marginalized out.

$$\begin{aligned} & P(x_i|\rho_i = c, \mathbf{x}_{\setminus i}) \\ &= \int P(x_i|\theta_c) P(\theta_c|\mathbf{x}_c) d\theta_c \\ &= \frac{\int F(x_i|\theta_c) \prod_{j \in c} F(x_j|\theta_c) H(\theta_c) d\theta_c}{\int \prod_{j \in c} F(x_j|\theta_c) H(\theta_c) d\theta_c} \end{aligned}$$



$$\begin{aligned} \rho|\alpha &\sim \mathrm{CRP}([n], \alpha) \\ \theta_c|H &\sim H \text{ for } c \in \rho \\ x_i|\theta_c &\sim F(\cdot|\theta_c) \text{ for } c \ni i \end{aligned}$$

- This is called a collapsed Gibbs sampler.

Random Trees in Bayesian Nonparametrics

Overview

- Bayesian nonparametric learning of trees and hierarchical partitions.

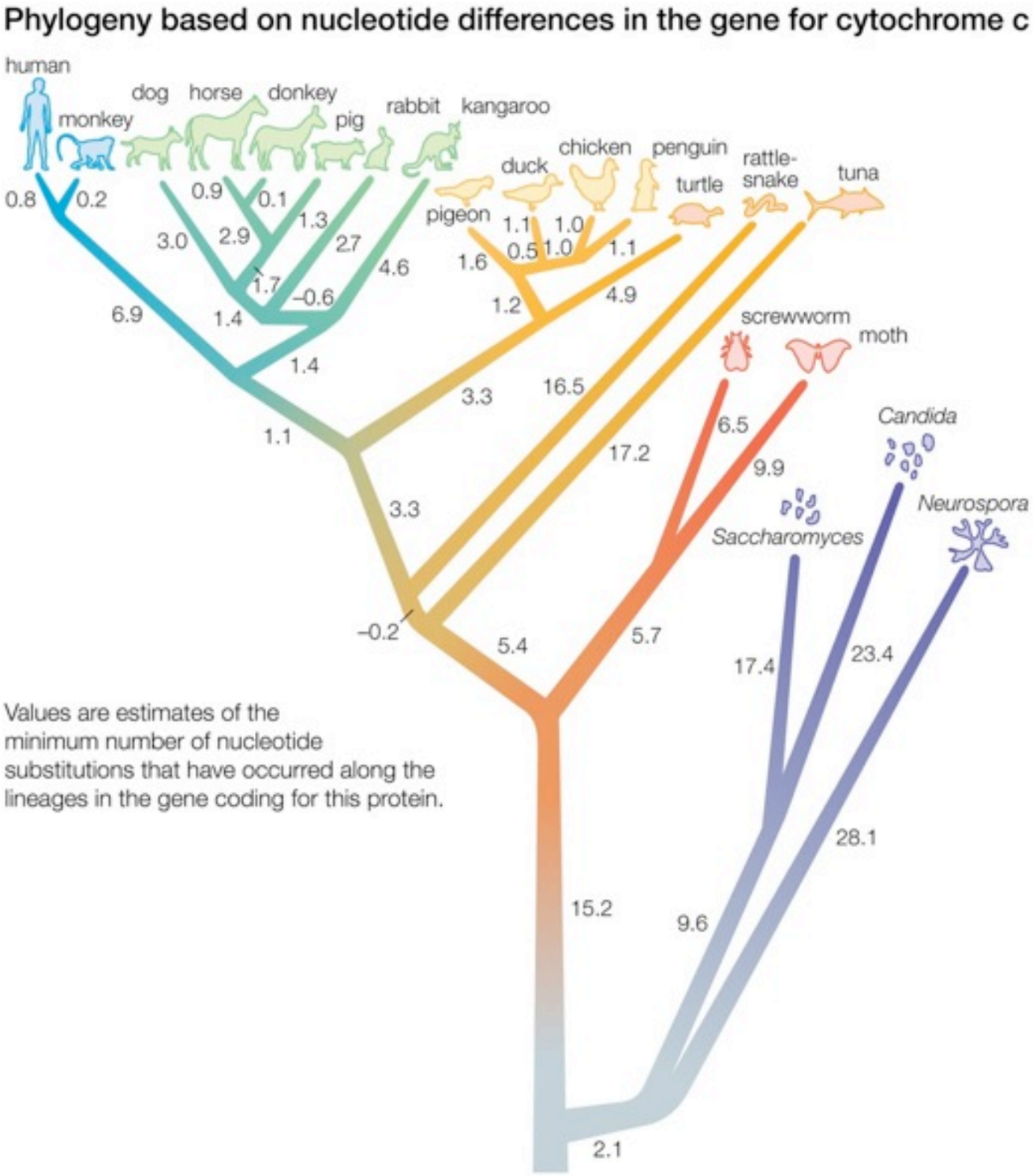
- View of rooted trees as sequences of partitions.

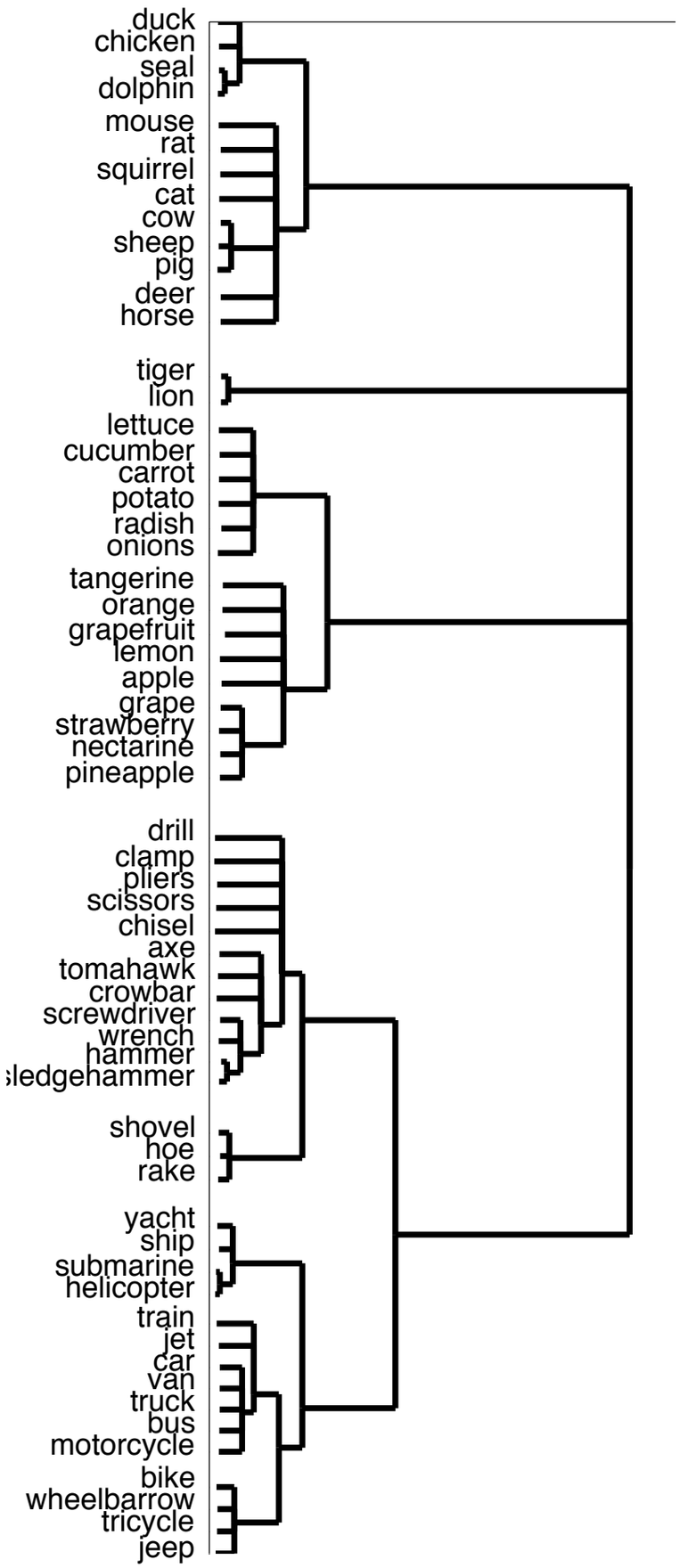
- Fragmentations and coagulations.

- 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

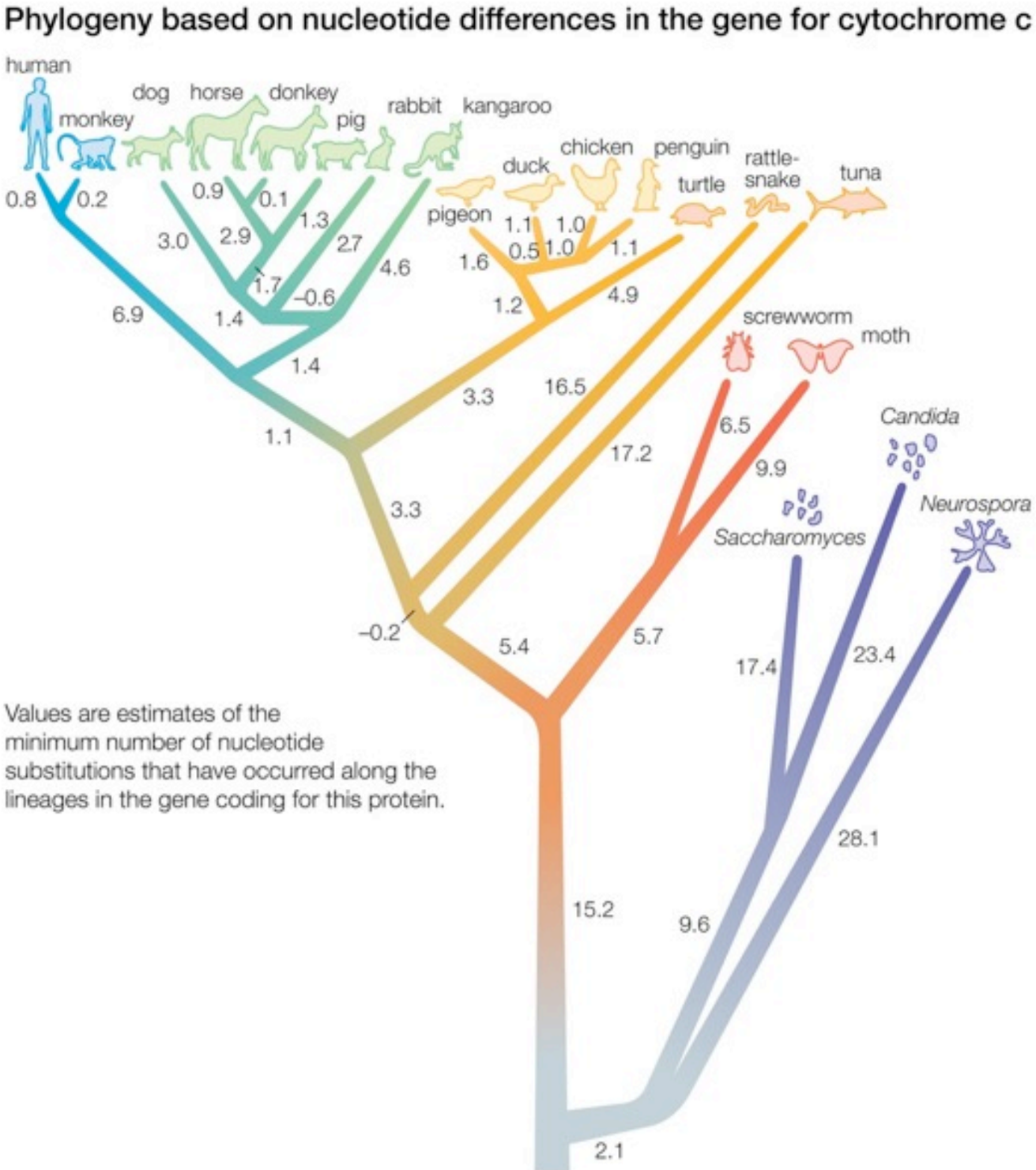
Bayesian Inference for Trees

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

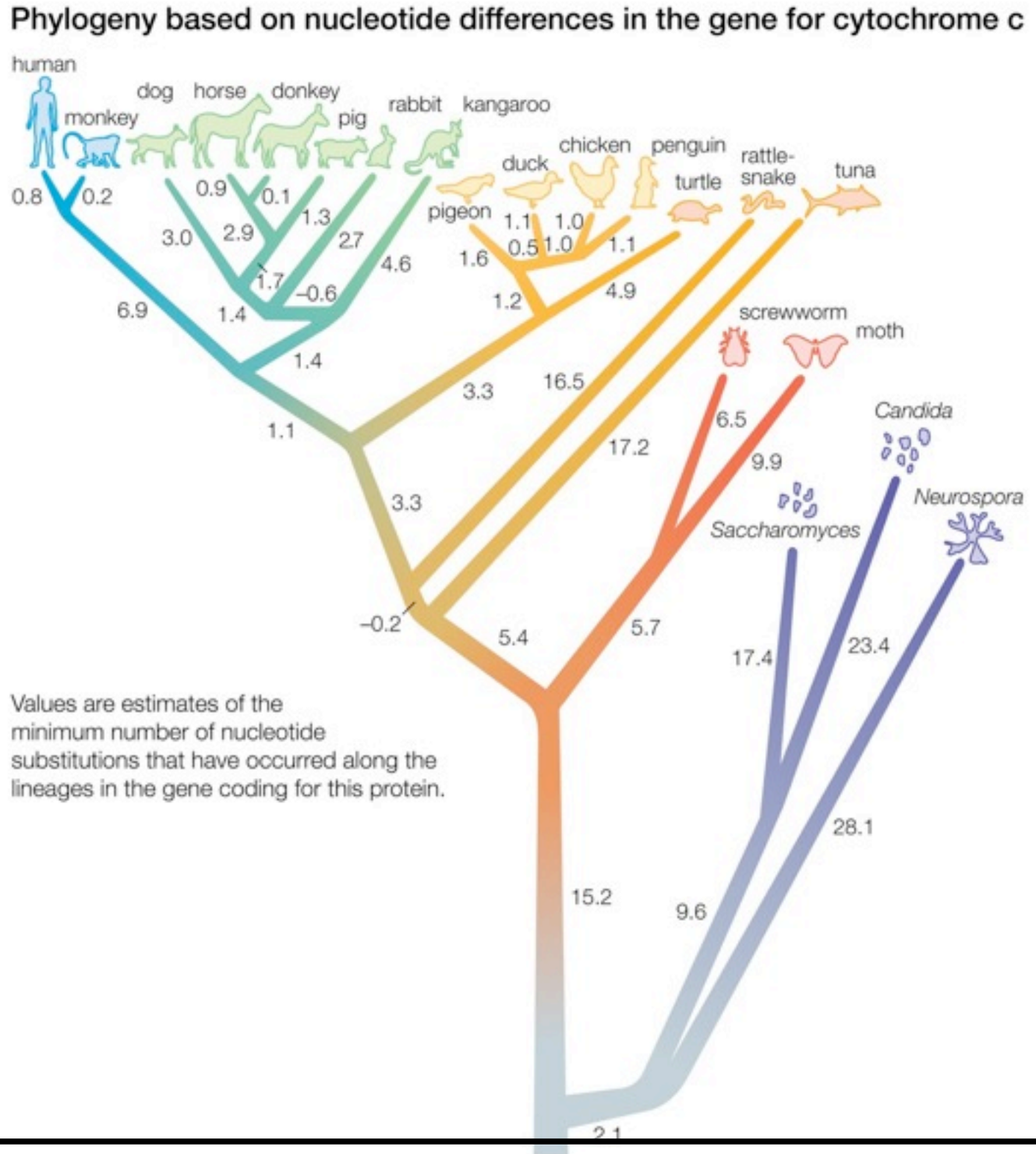
$$P(T|\mathbf{x}) \propto P(T)P(\mathbf{x}|T)$$

- Projectivity and exchangeability leads to Bayesian nonparametric models for P(T).

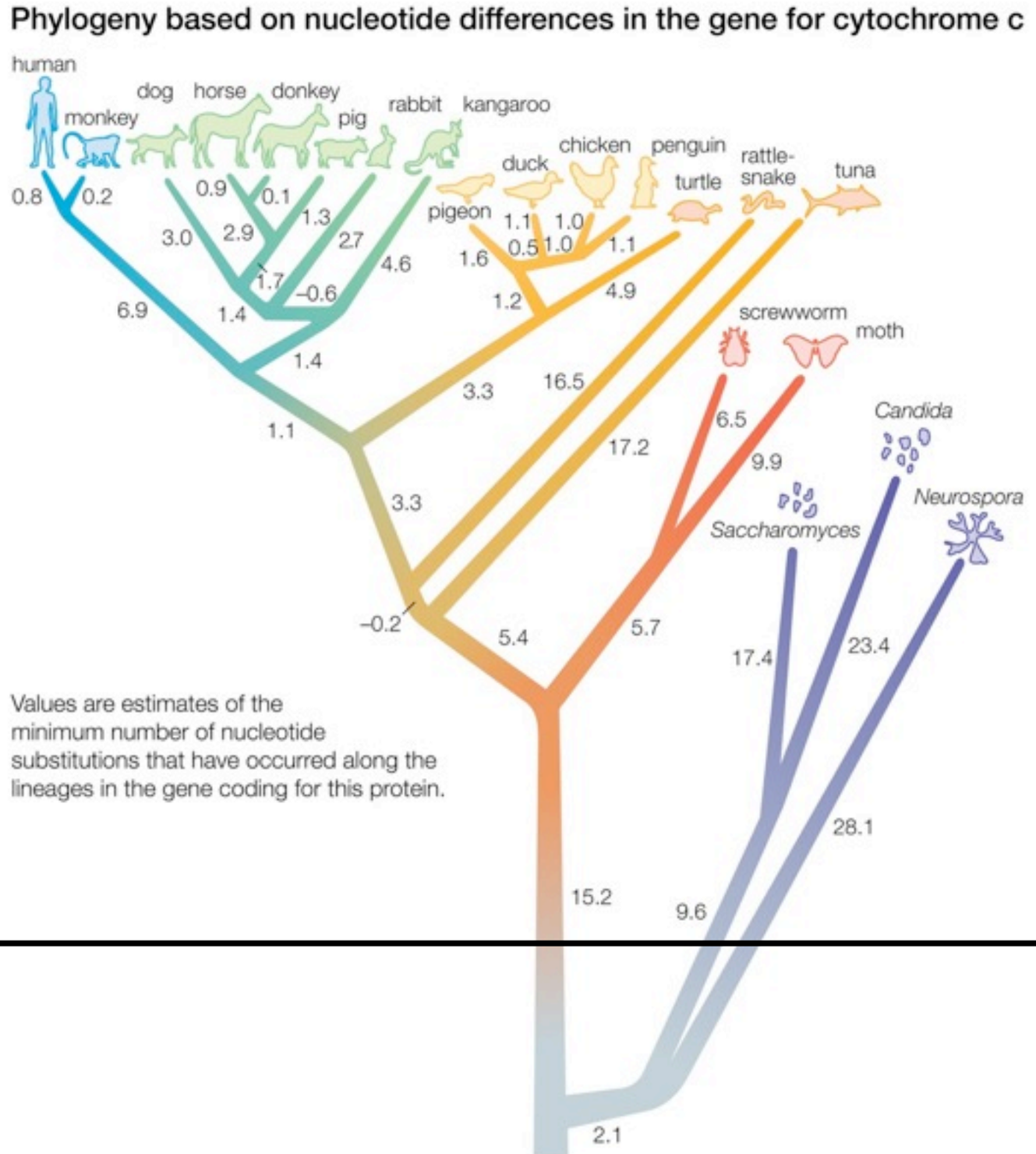
Trees as Sequences of Partitions



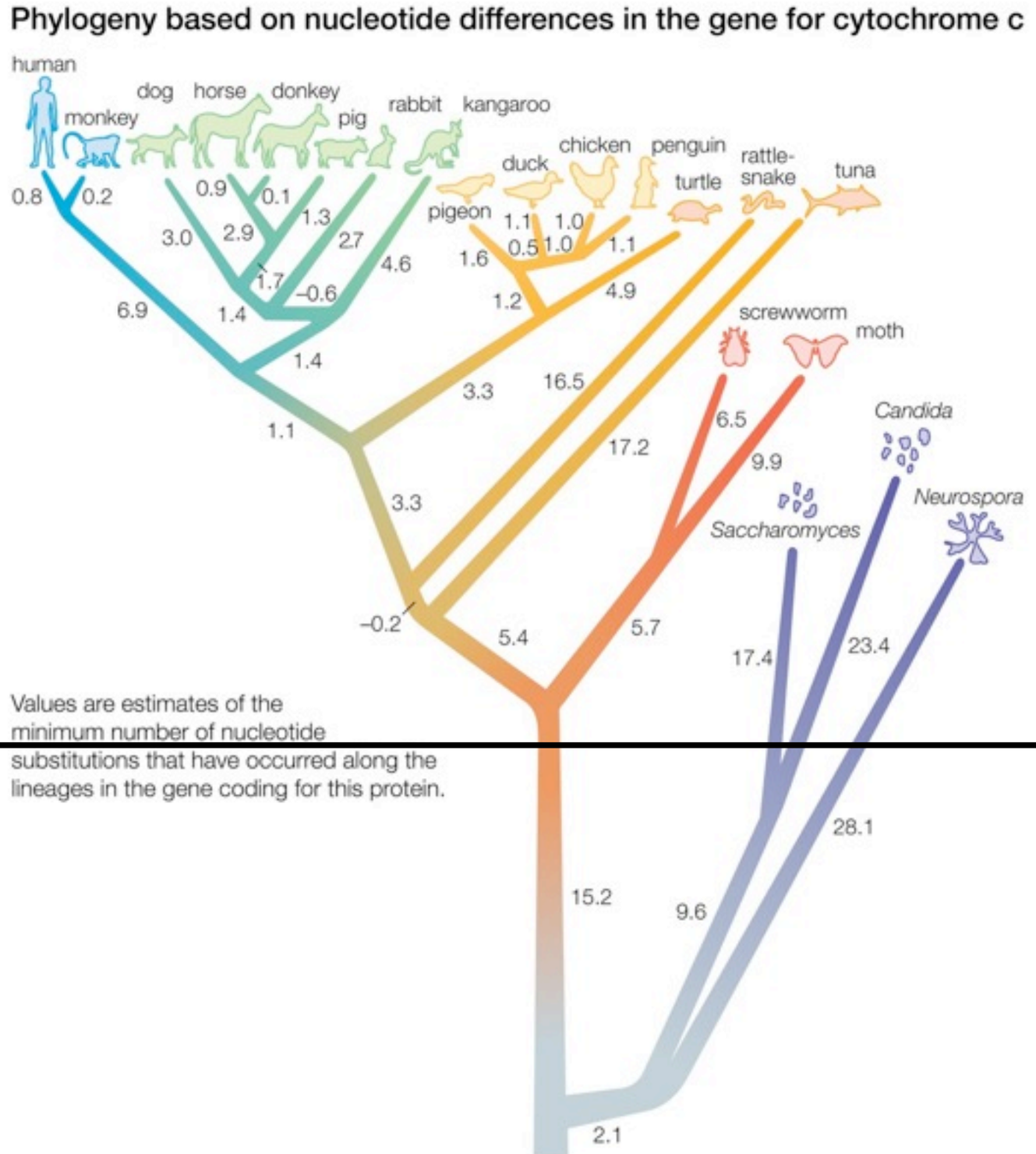
Trees as Sequences of Partitions



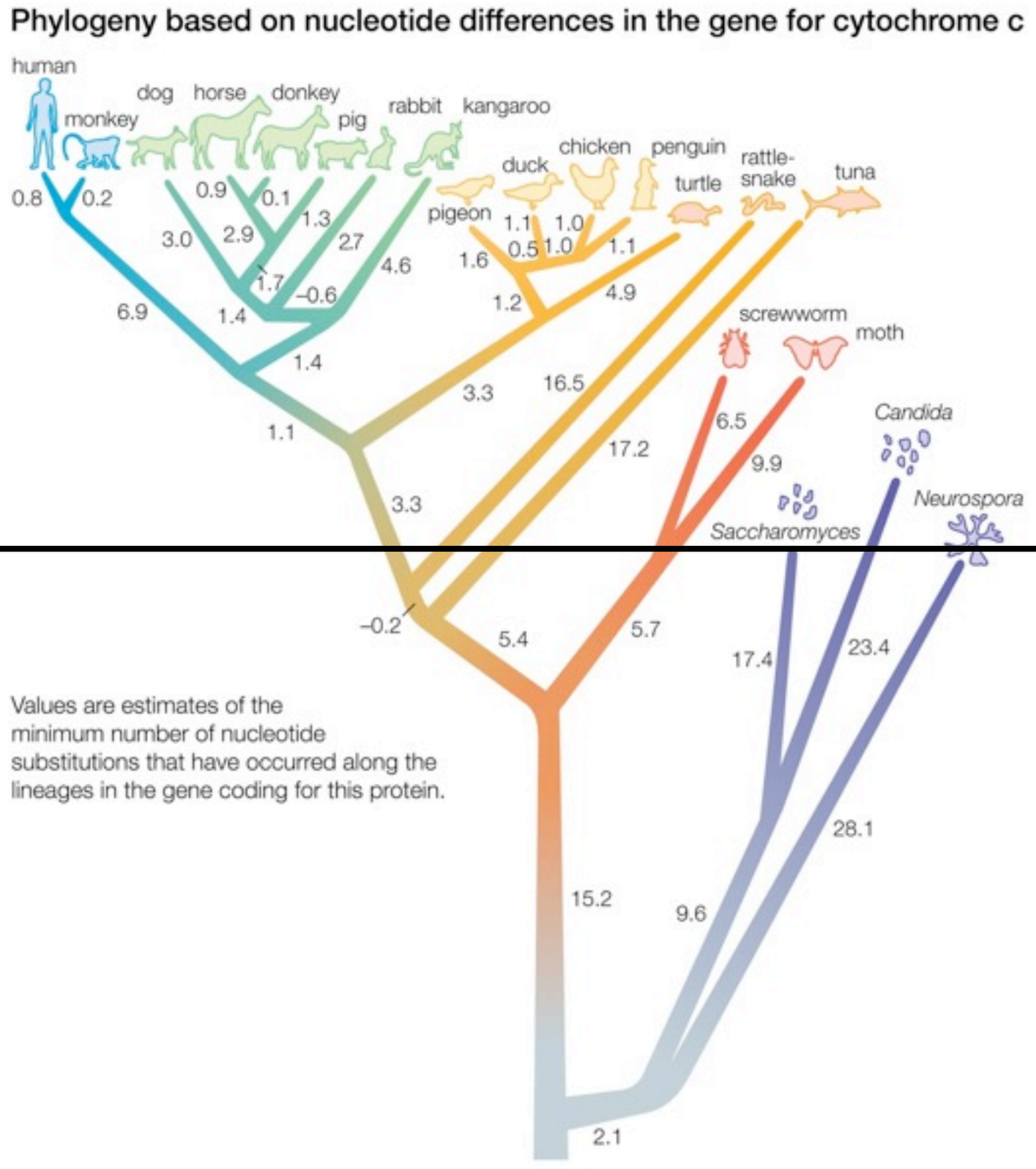
Trees as Sequences of Partitions



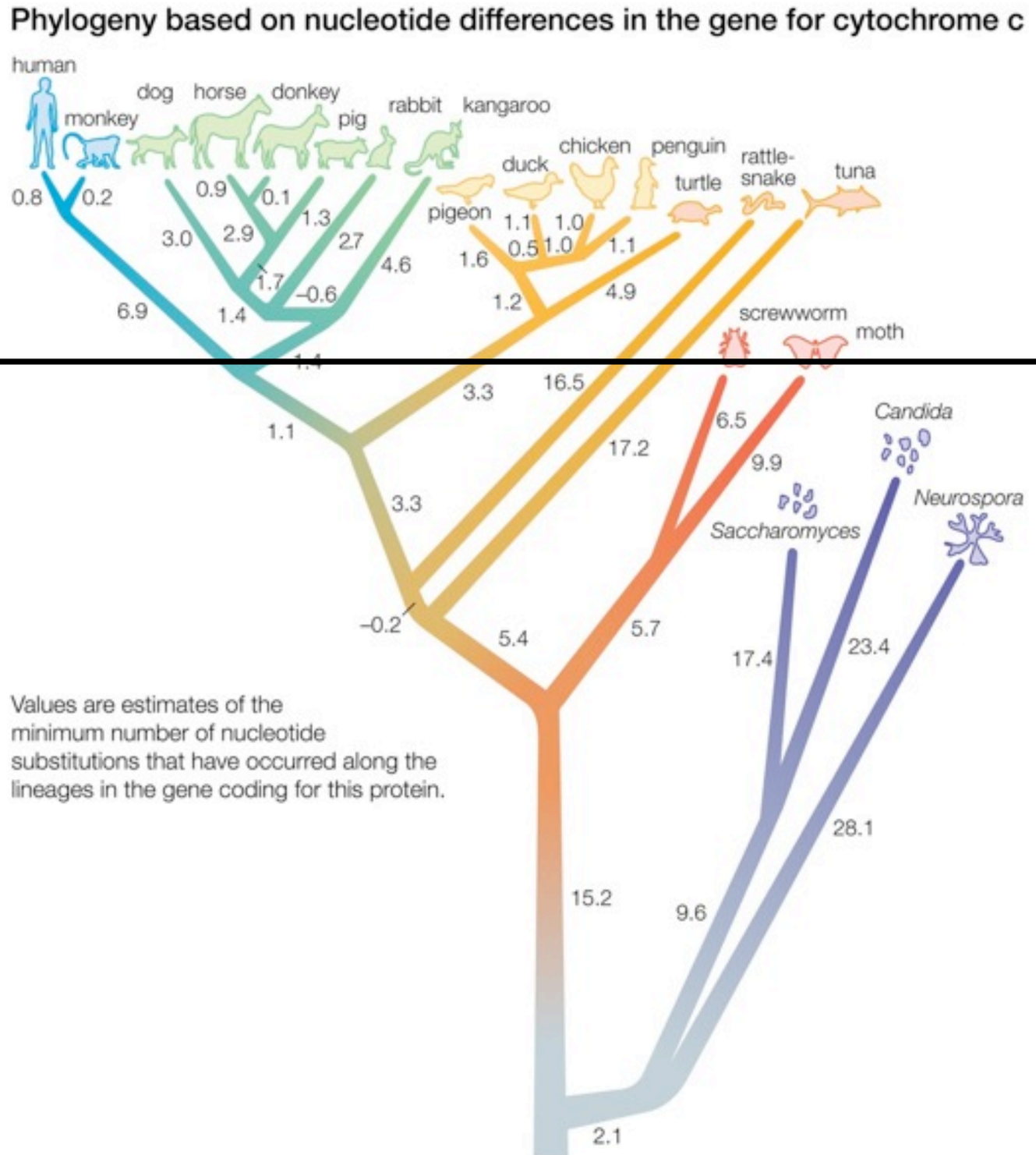
Trees as Sequences of Partitions



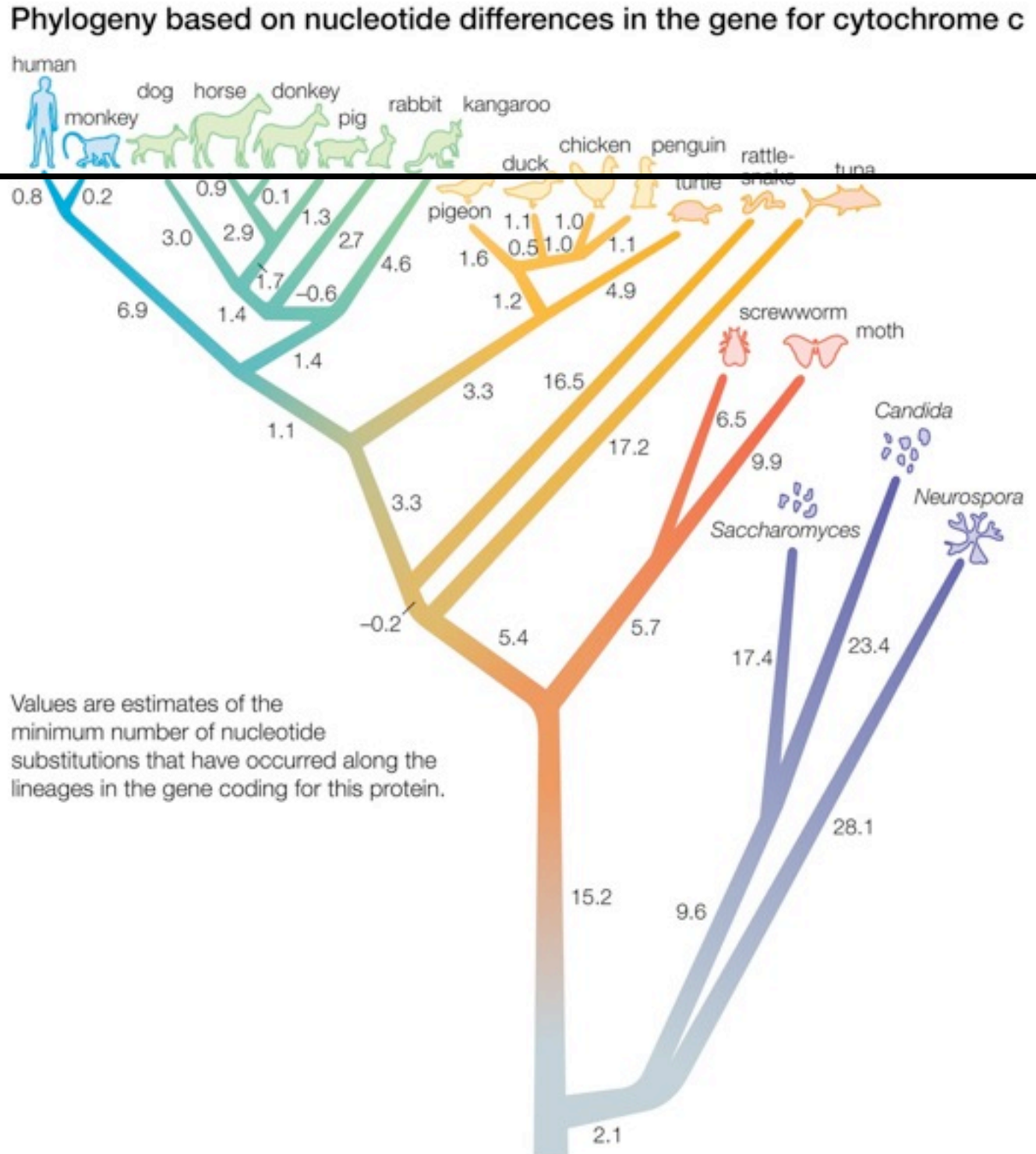
Trees as Sequences of Partitions



Trees as Sequences of Partitions



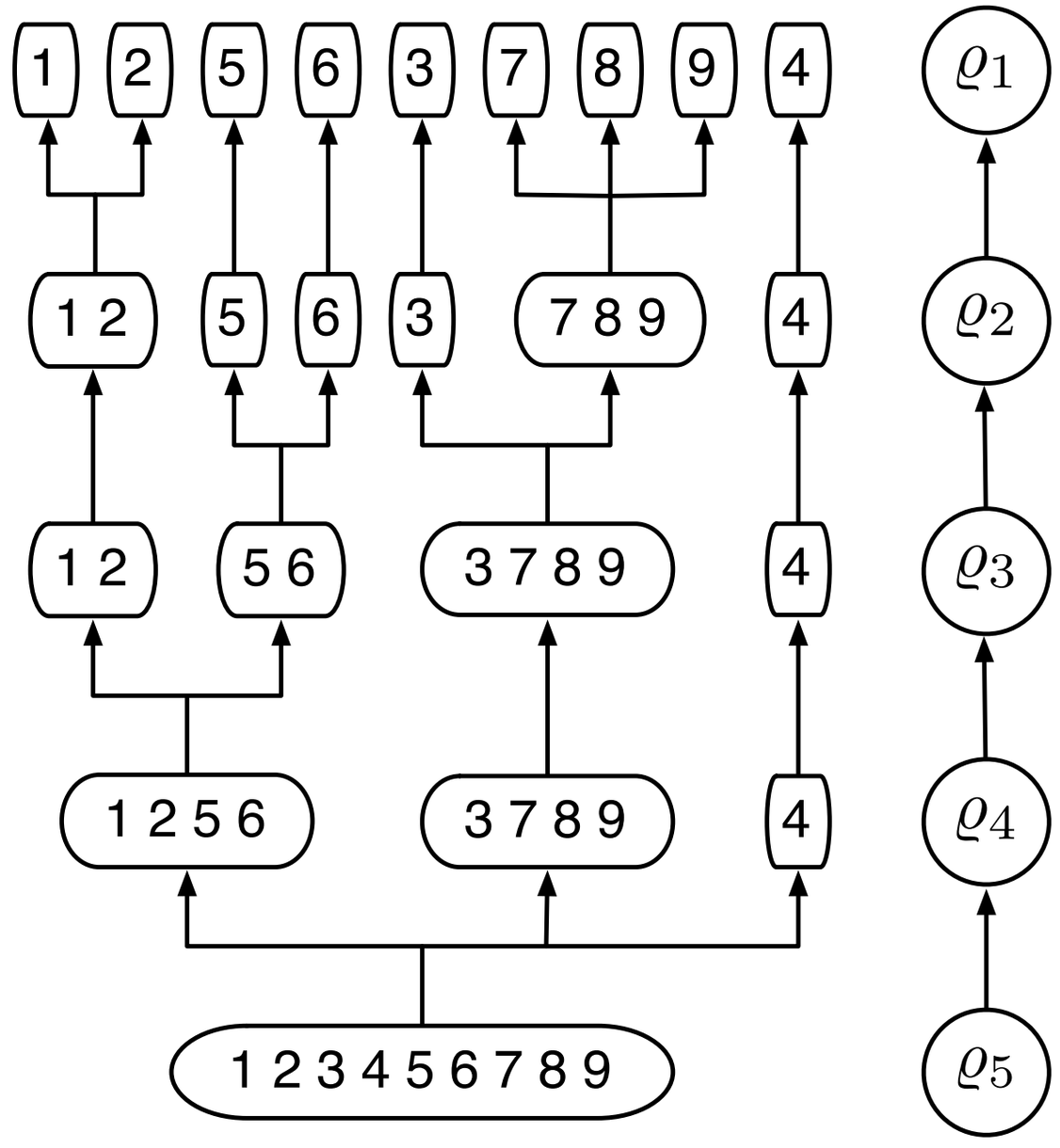
Trees as Sequences of Partitions



Fragmenting Partitions

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

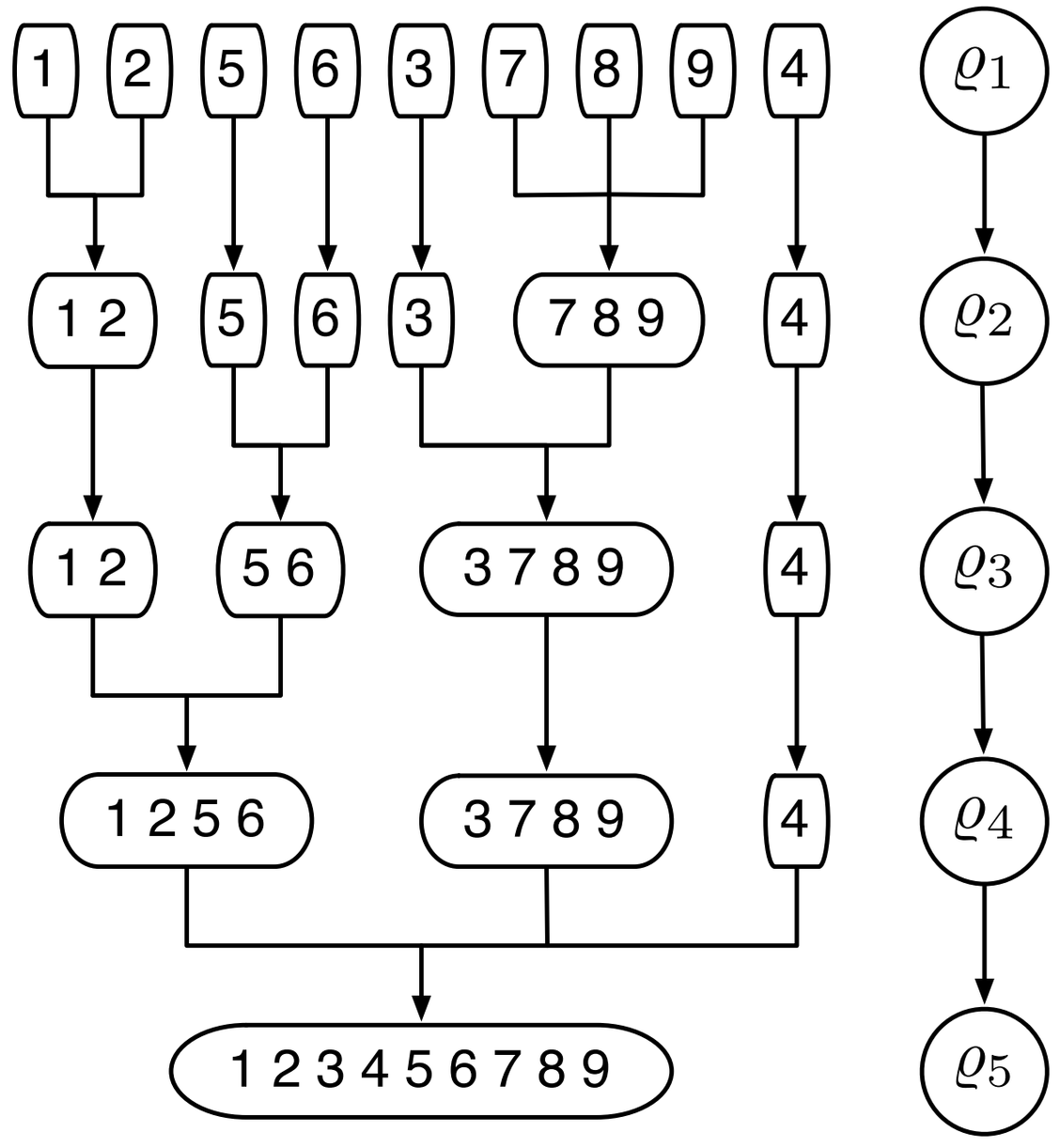
- *Can define a distribution over trees using a Markov chain of fragmenting partitions, with absorbing state* $\mathbf{0}_S$ *(partition where all data items are in their own clusters).*



Coagulating Partitions

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

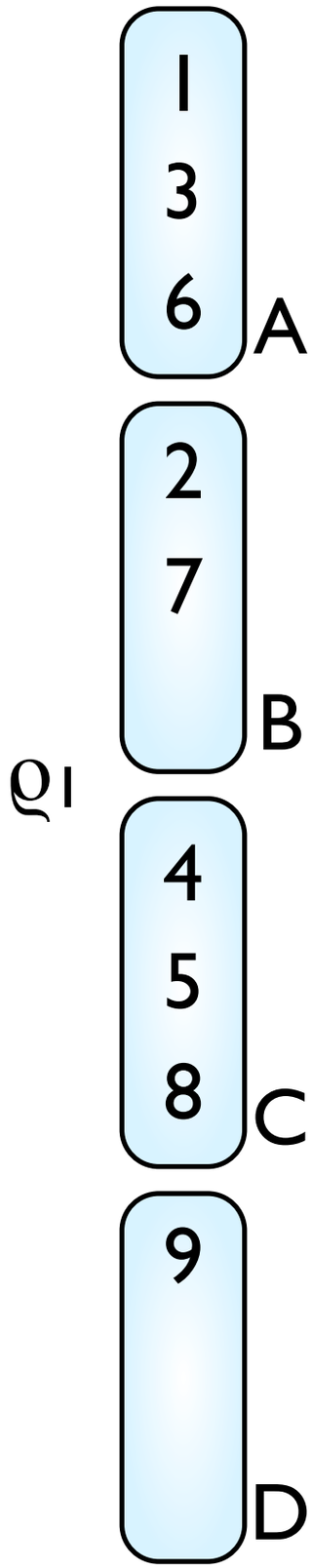
- *Can define a distribution over trees by using a Markov chain of coagulating partitions, with absorbing state* $\mathbf{1}_S$ *(partition where all data items are in one cluster).*



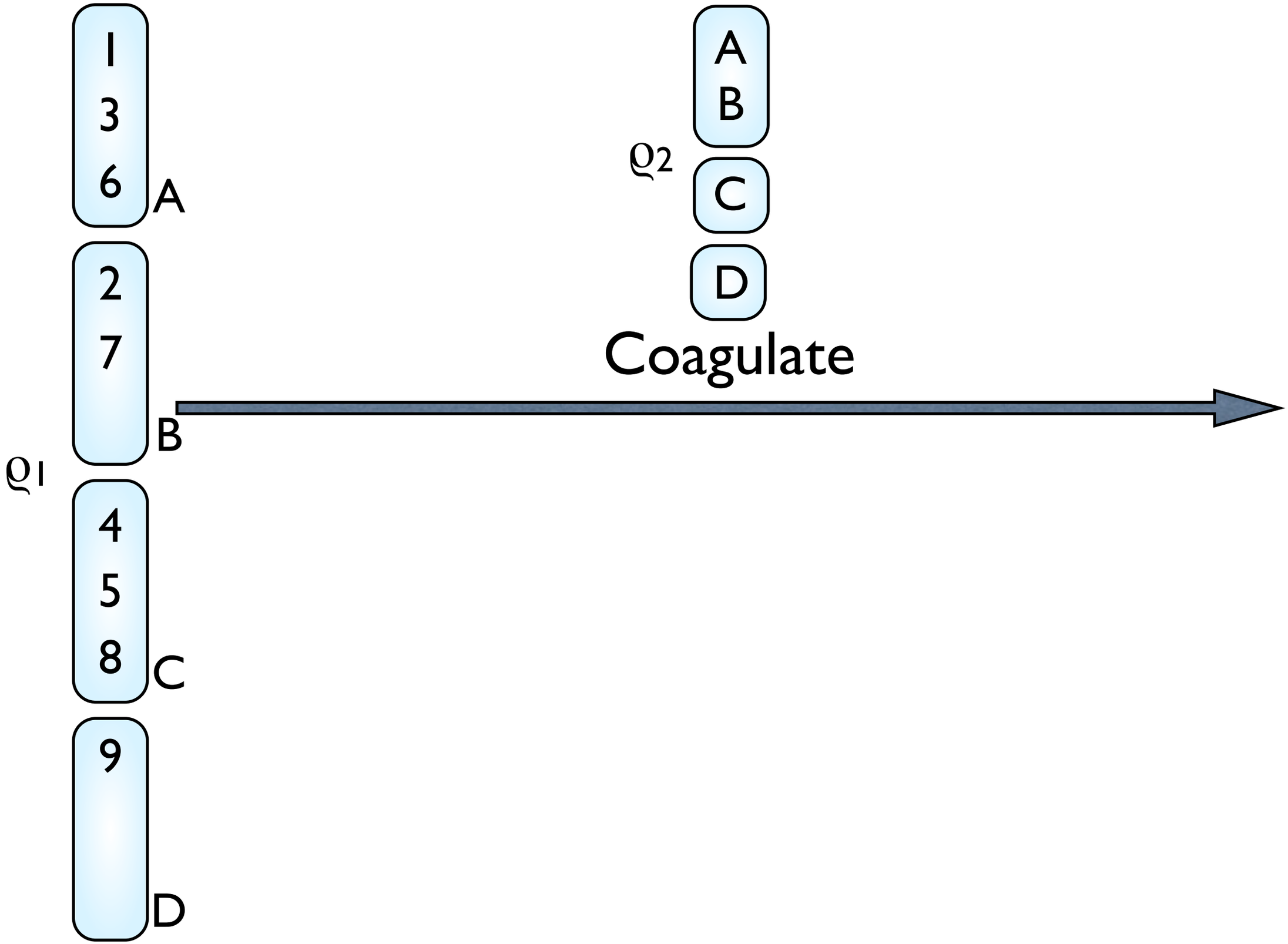
Random Fragmentations and Random Coagulations

[Bertoin 2006]

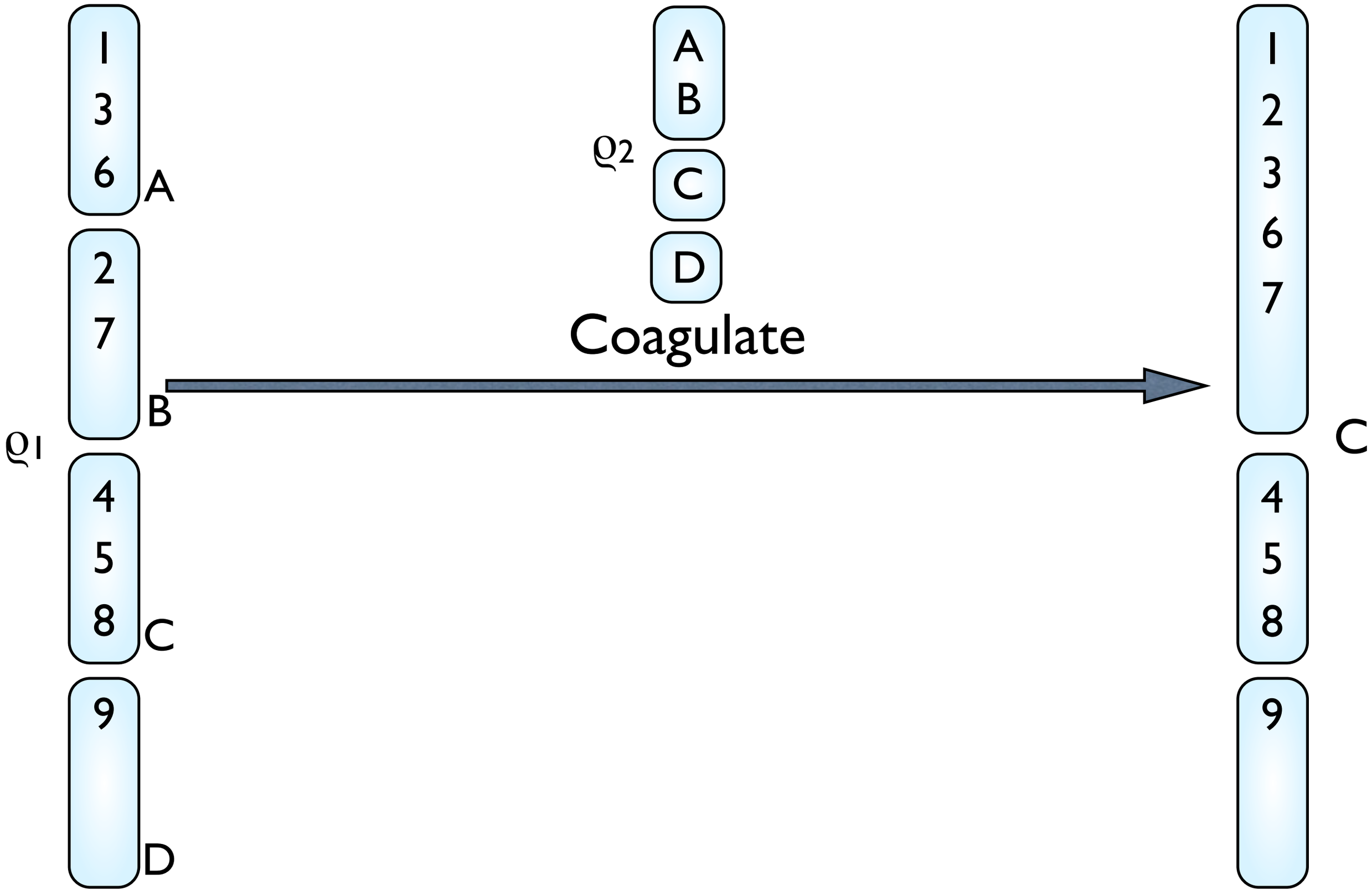
Coagulation and Fragmentation Operators



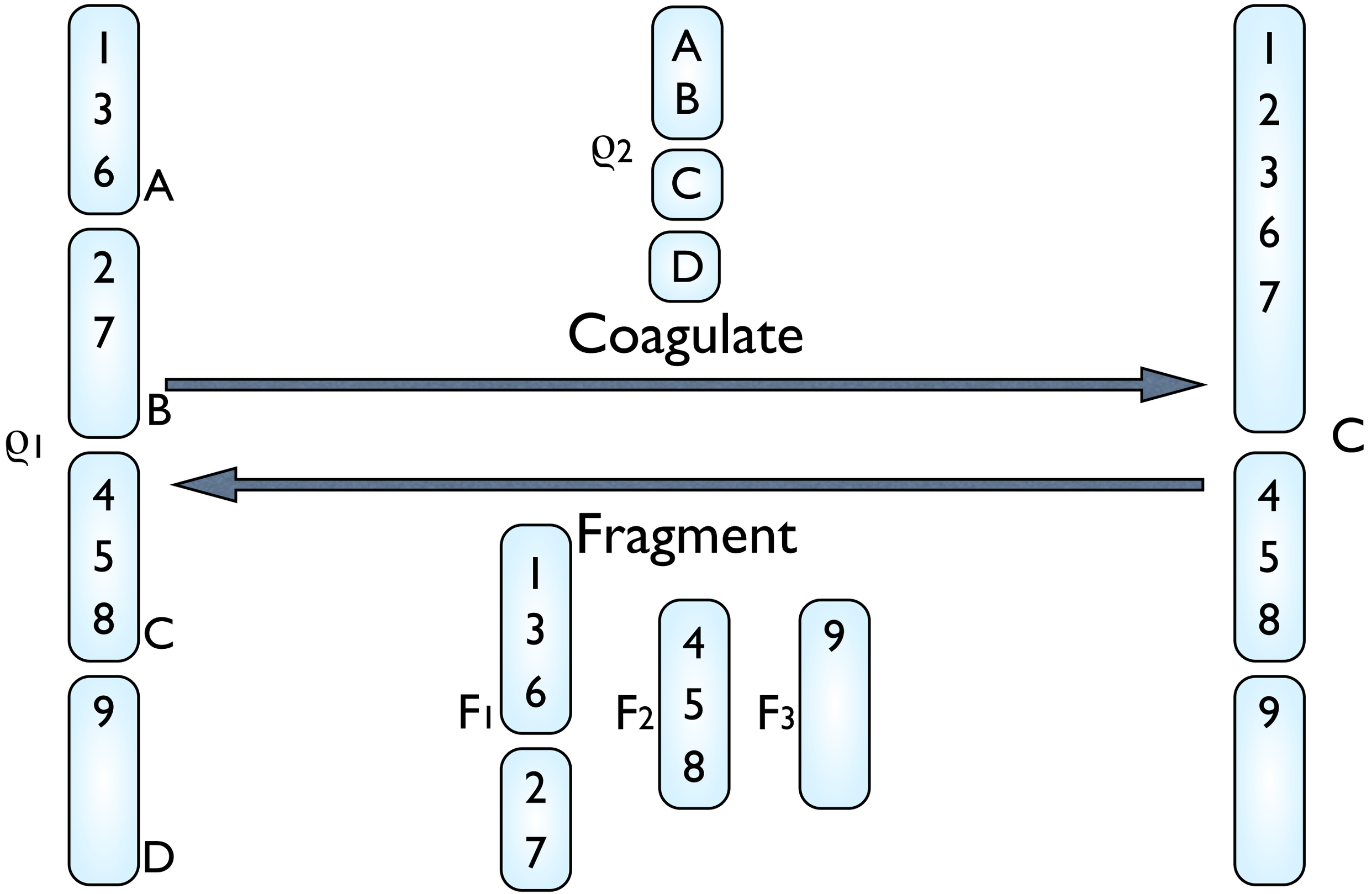
Coagulation and Fragmentation Operators



Coagulation and Fragmentation Operators



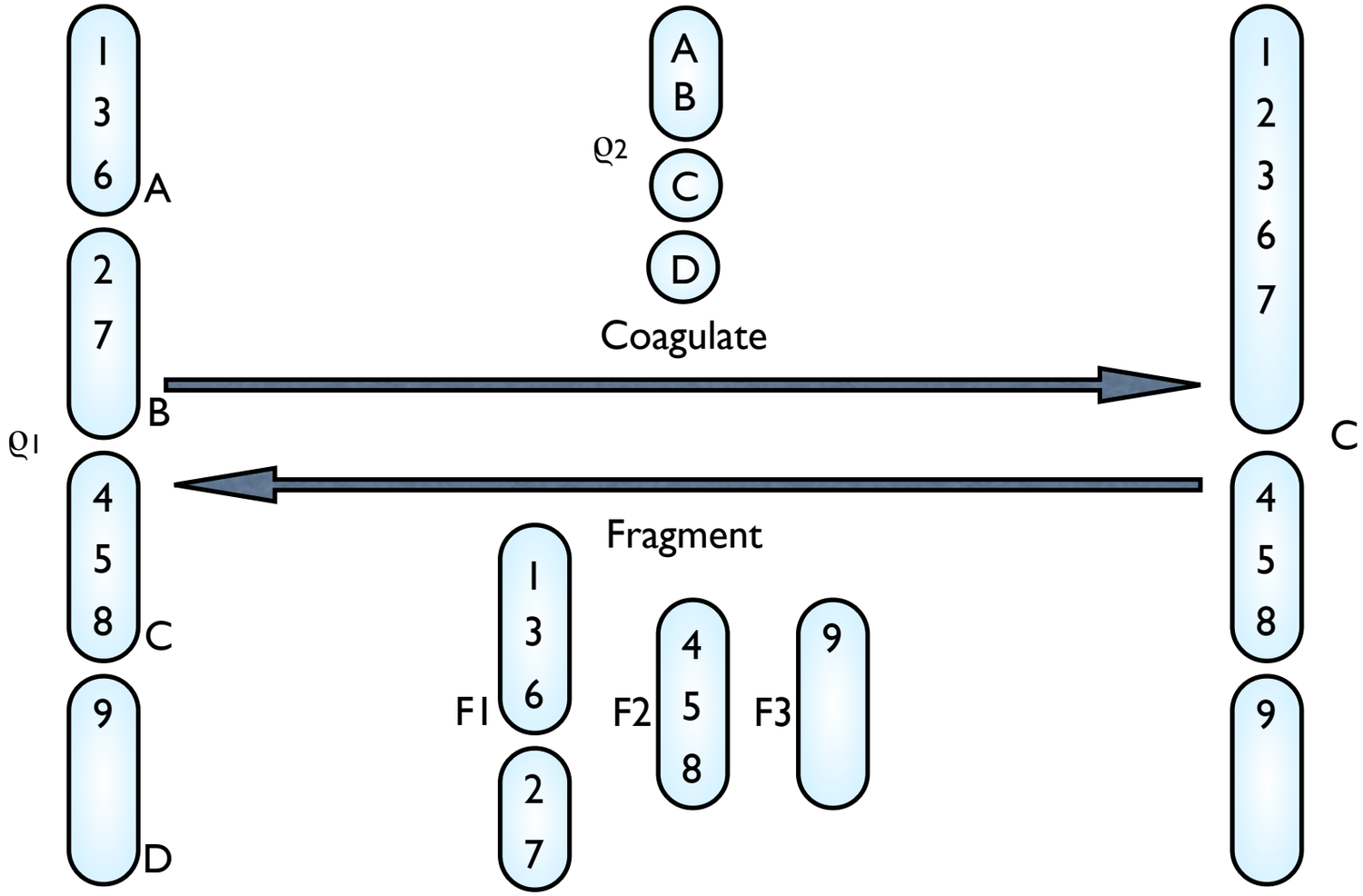
Coagulation and Fragmentation Operators



Random 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

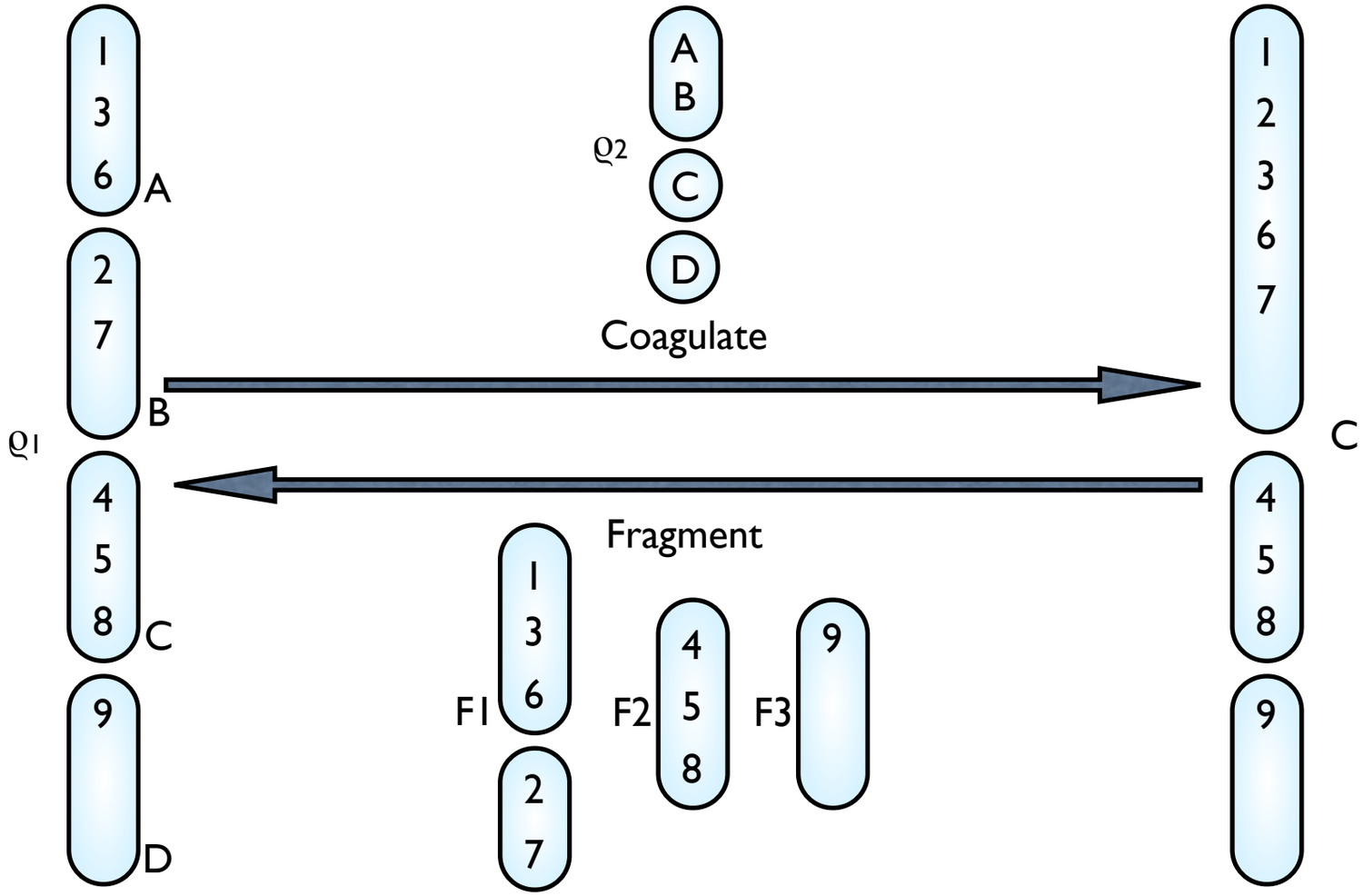
$$\varrho_2 \mid \varrho_1 \sim \mathrm{CRP}(\varrho_1, d, \alpha).$$



Random Fragmentations

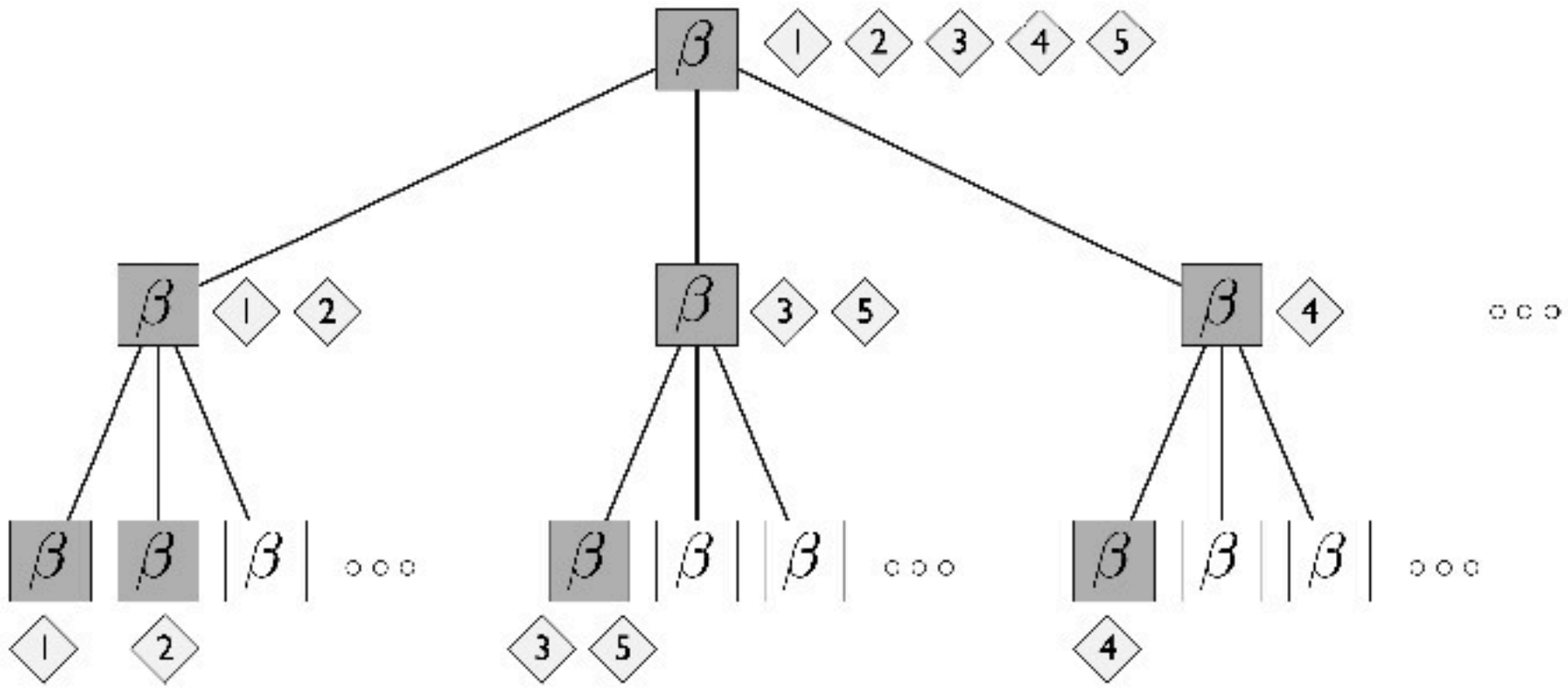
- Let $C \in \mathcal{P}_{[n]}$ and for each $c \in C$ let $F_c \in \mathcal{P}_c$.
  - Denote **fragmentation** of C by $\{F_c\}$ as $\mathrm{frag}(C,\{F_c\})$.
  - Write $\varrho_1 \mid C \sim \mathrm{FRAG}(C,d,\alpha)$ if $\varrho_1 = \mathrm{frag}(C,\{F_c\})$ with

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



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.

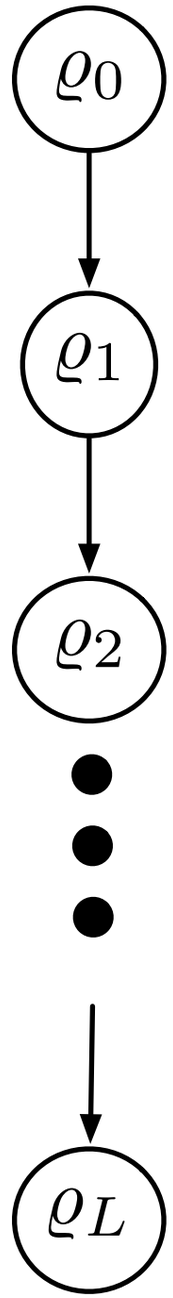
[Blei et al NIPS 2004, JACM 2010]

Nested Chinese Restaurant Processes

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

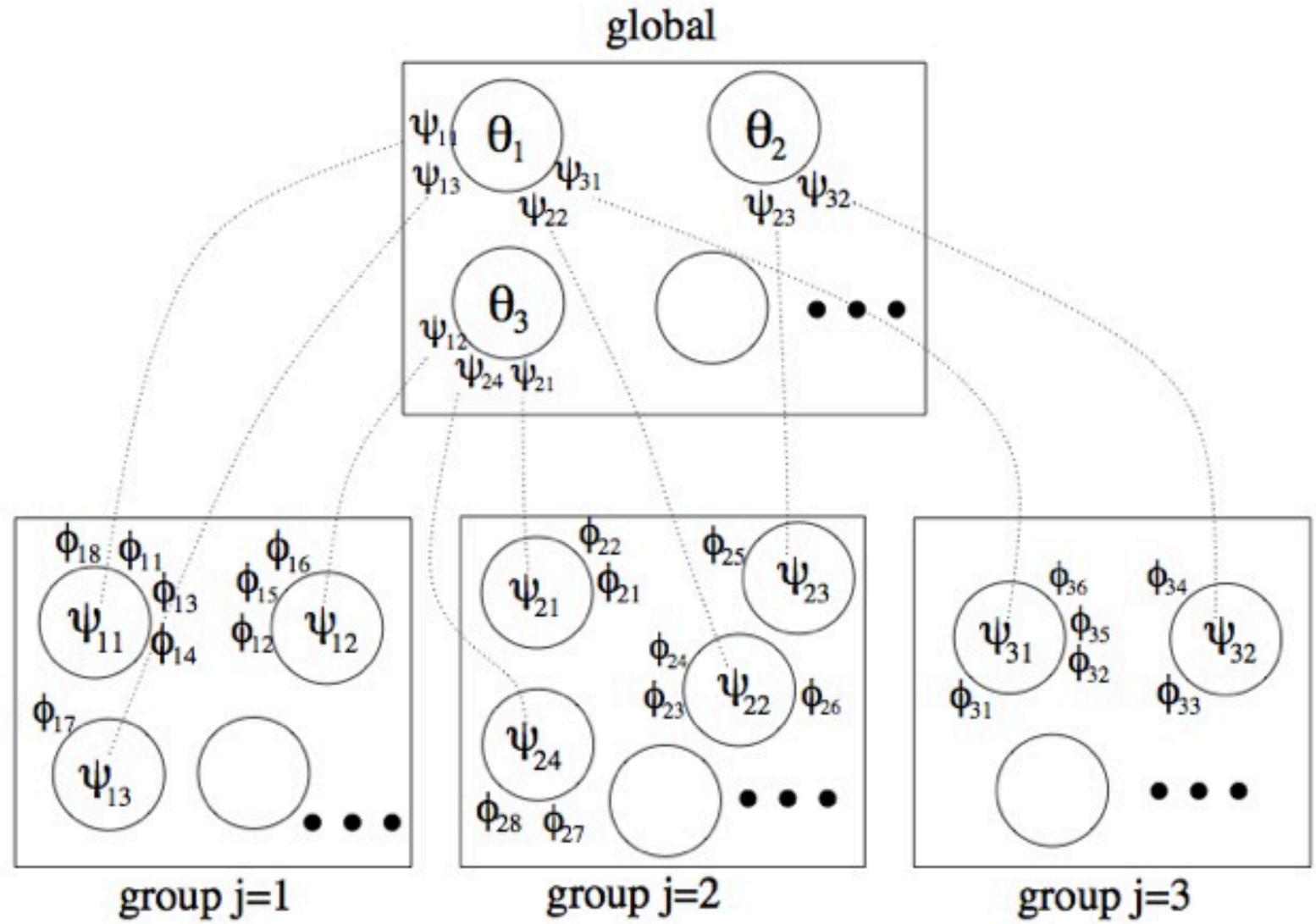
$$\varrho_l = \text{FRAG}(\varrho_{l-1}, 0, \alpha_l)$$

- 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).



[Rodriguez et al JASA 2008]

Chinese Restaurant Franchise



The metaphor is as follows (see Fig. 2). We have a restaurant franchise with a shared menu across the restaurants. At each table of each restaurant, one dish is ordered from the menu by the first customer who sits there, and this dish is shared among all of the customers who sit at that table. Multiple tables in multiple restaurants can serve the same dish.

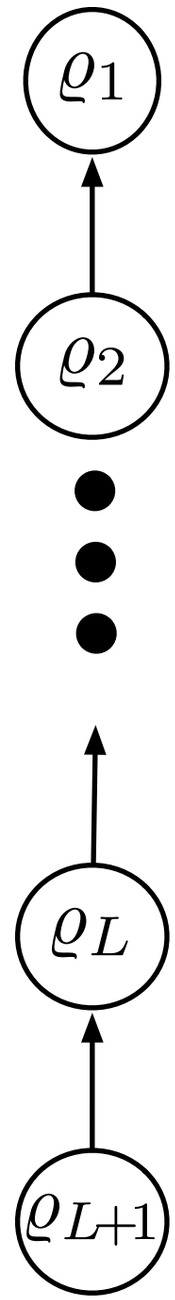
[Teh et al JASA 2006]

Chinese Restaurant Franchise

- For a simple linear hierarchy (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 $\varrho_{L+1} = \{\{1\},\{2\},\ldots,\{n\}\}$.
  - For each level $l = L,L-1,\ldots,1$:

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

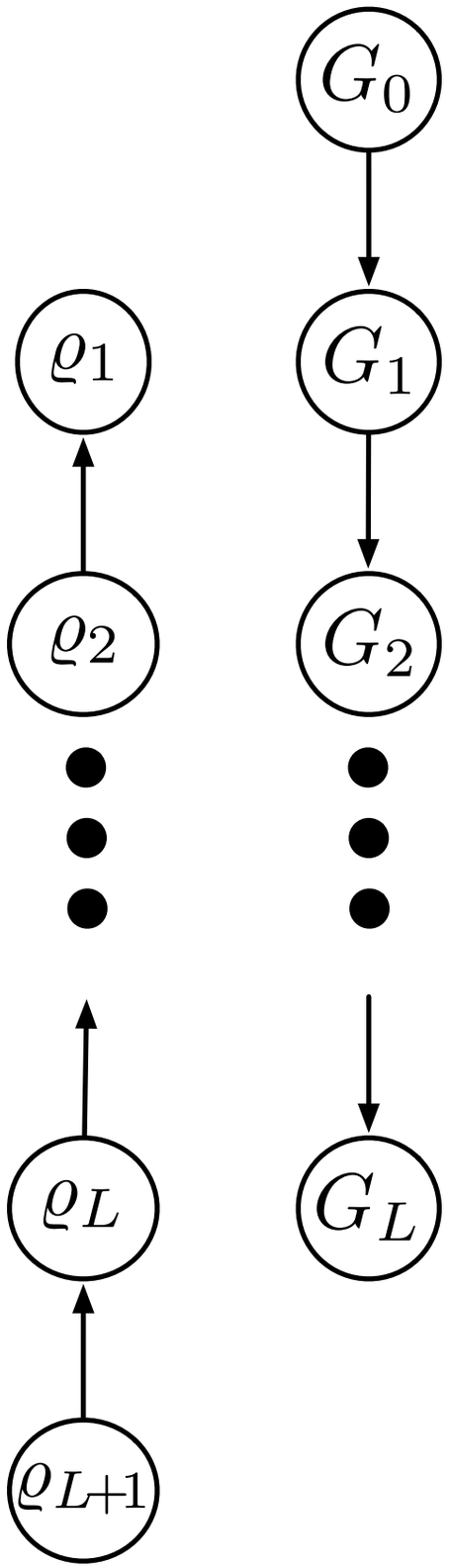
- This is also Markov chain of partitions.



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).

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



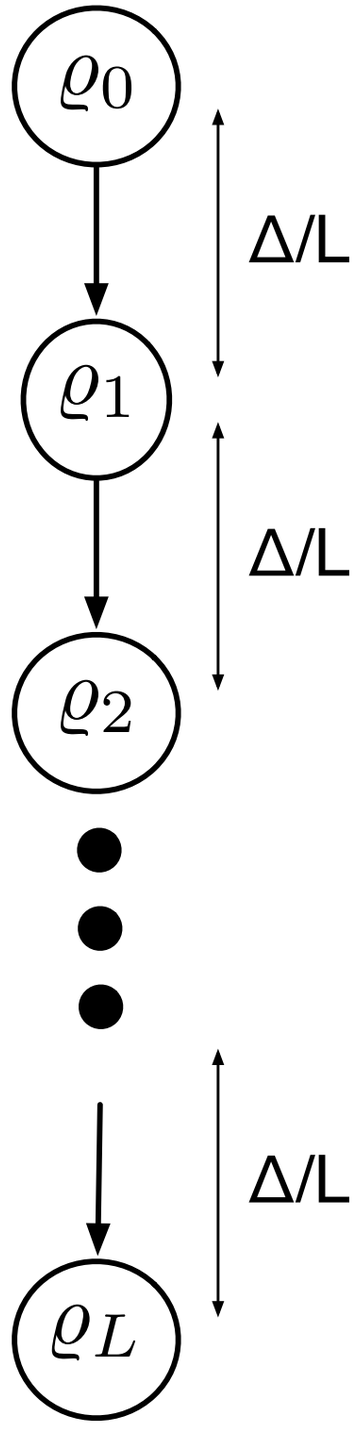
- Generalizable to tree-structured hierarchies and **hierarchical Pitman-Yor processes**.
- The CRF has been rarely used as a model of hierarchical partitions. Typically it is only used as a convenient representation for inference in the HDP and HPYP.

[Teh et al JASA 2006, Goldwater et al NIPS 2006, Teh ACL 2006]

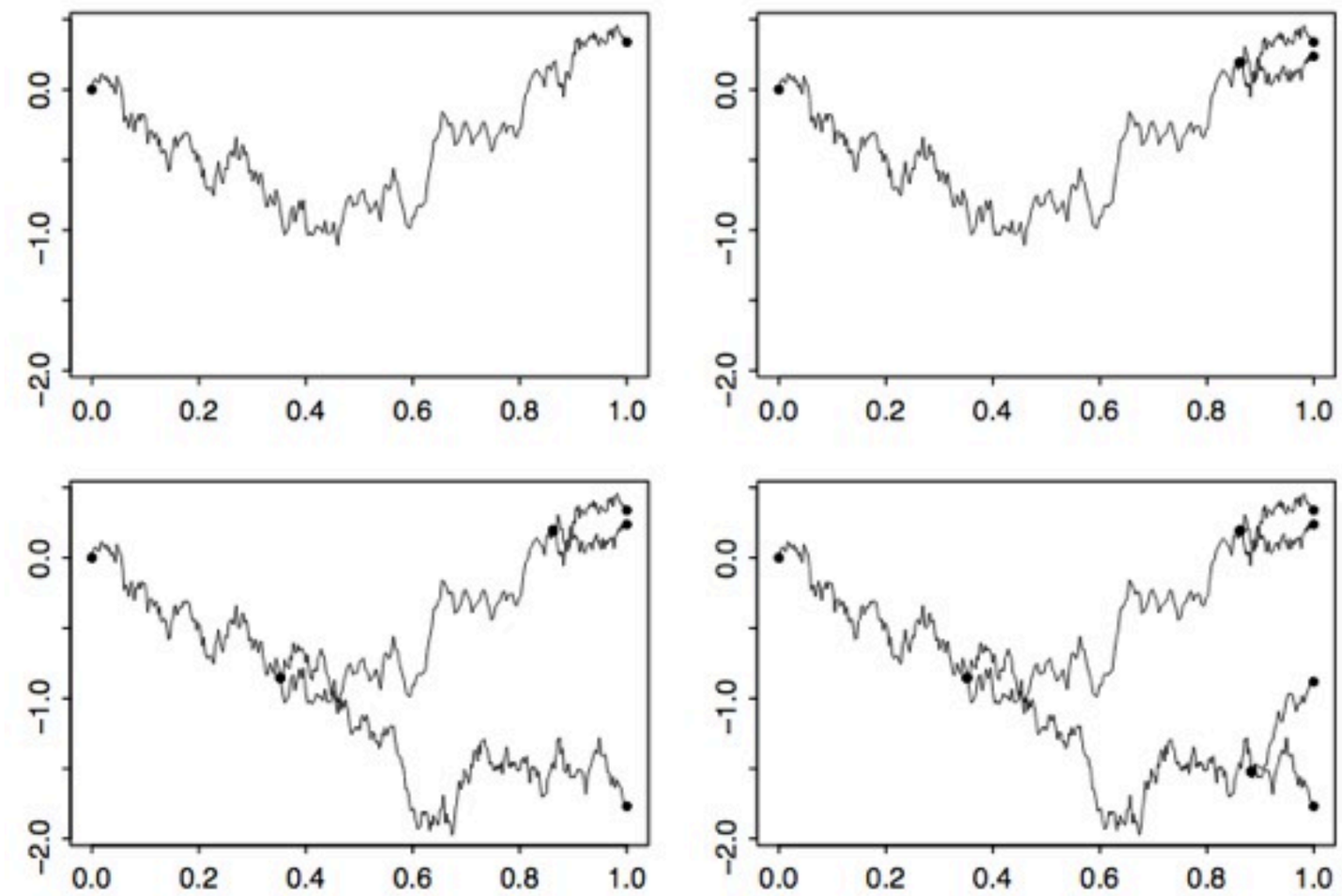
Continuum Limit of Partition-valued Markov Chains

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 BA 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**.

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 $\varrho_0 = \{\{1\},\{2\},\ldots,\{n\}\}$.
  - 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**.

[Kingman SP&A 1982, JoAP 1982]

Kingman's Coalescent

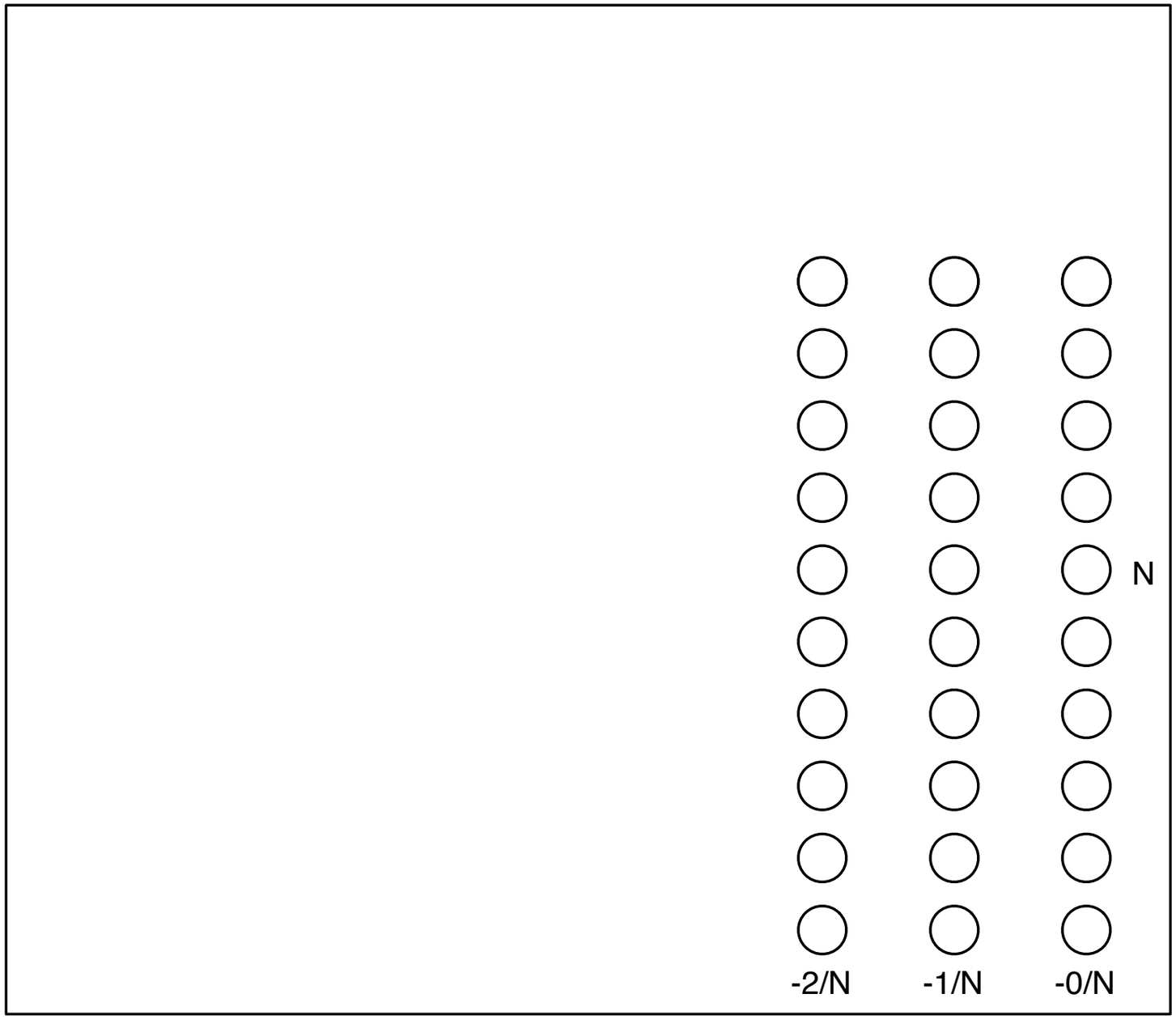
- Derived from the Wright-Fisher model of population genetics.
- Model of the genealogies of n haploid individuals among a size N population.
- Gives a tree-structured genealogy because each individual assumed to have one parent.

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.

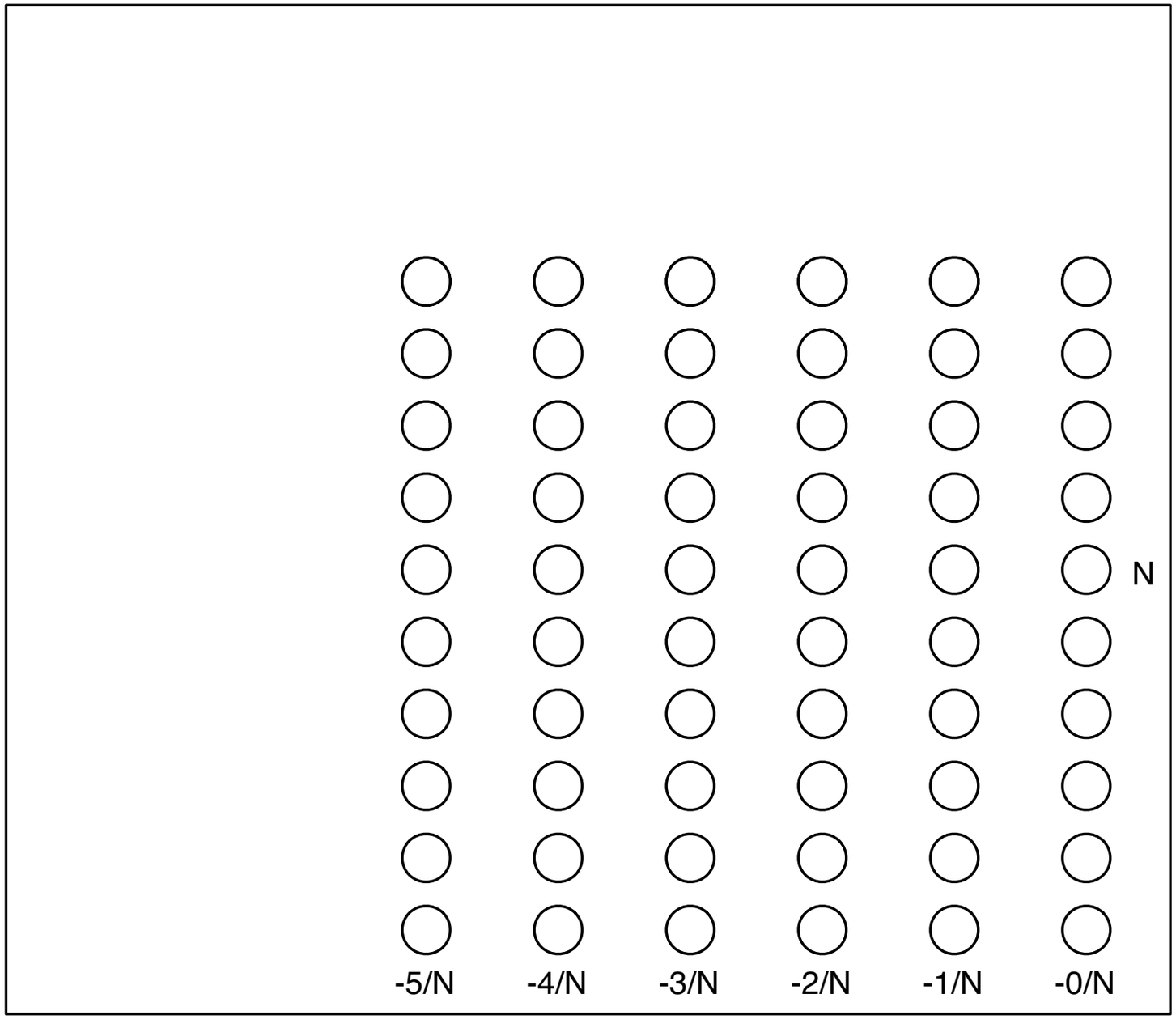
Kingman's Coalescent

- 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.



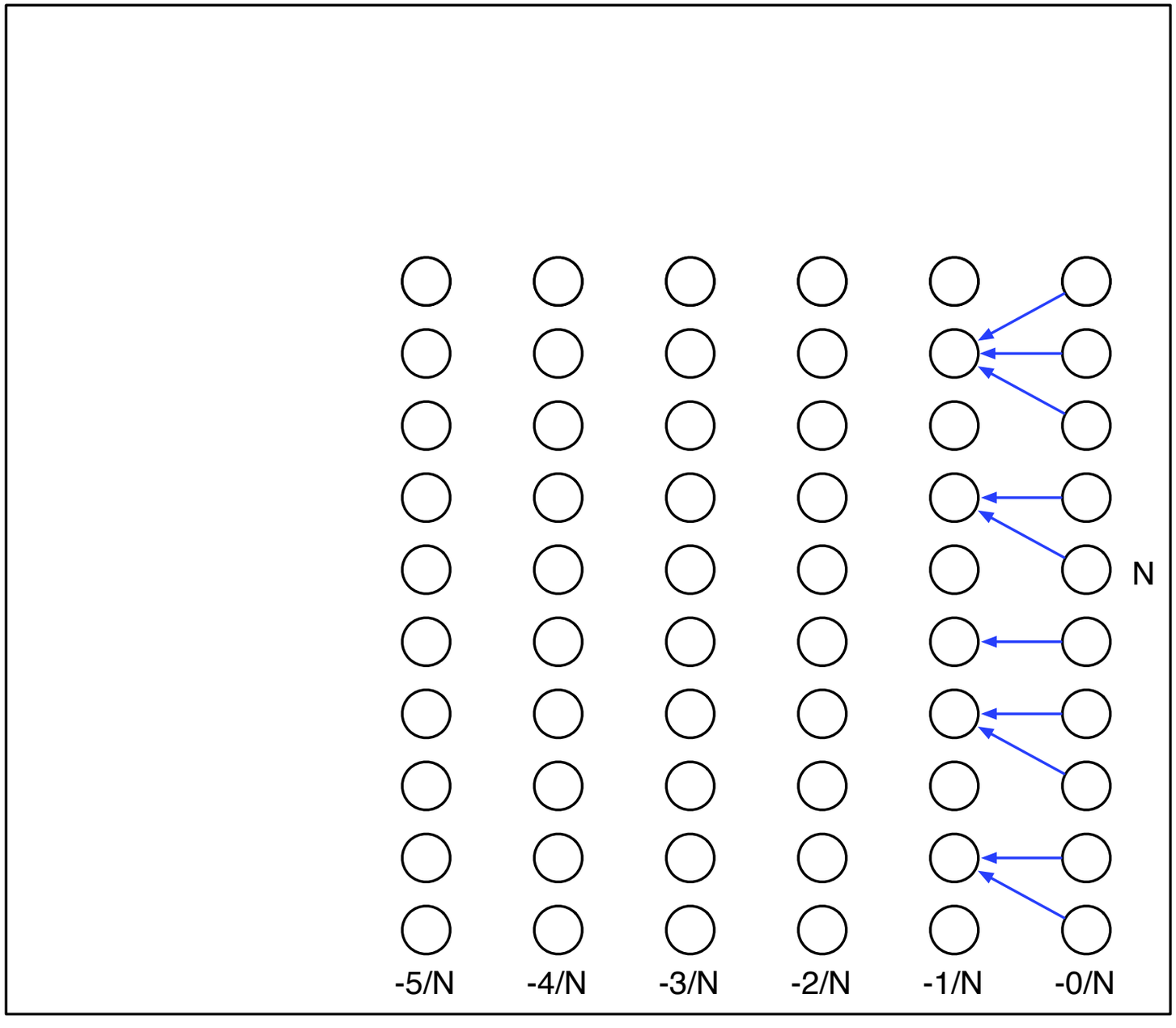
Kingman's Coalescent

- 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.



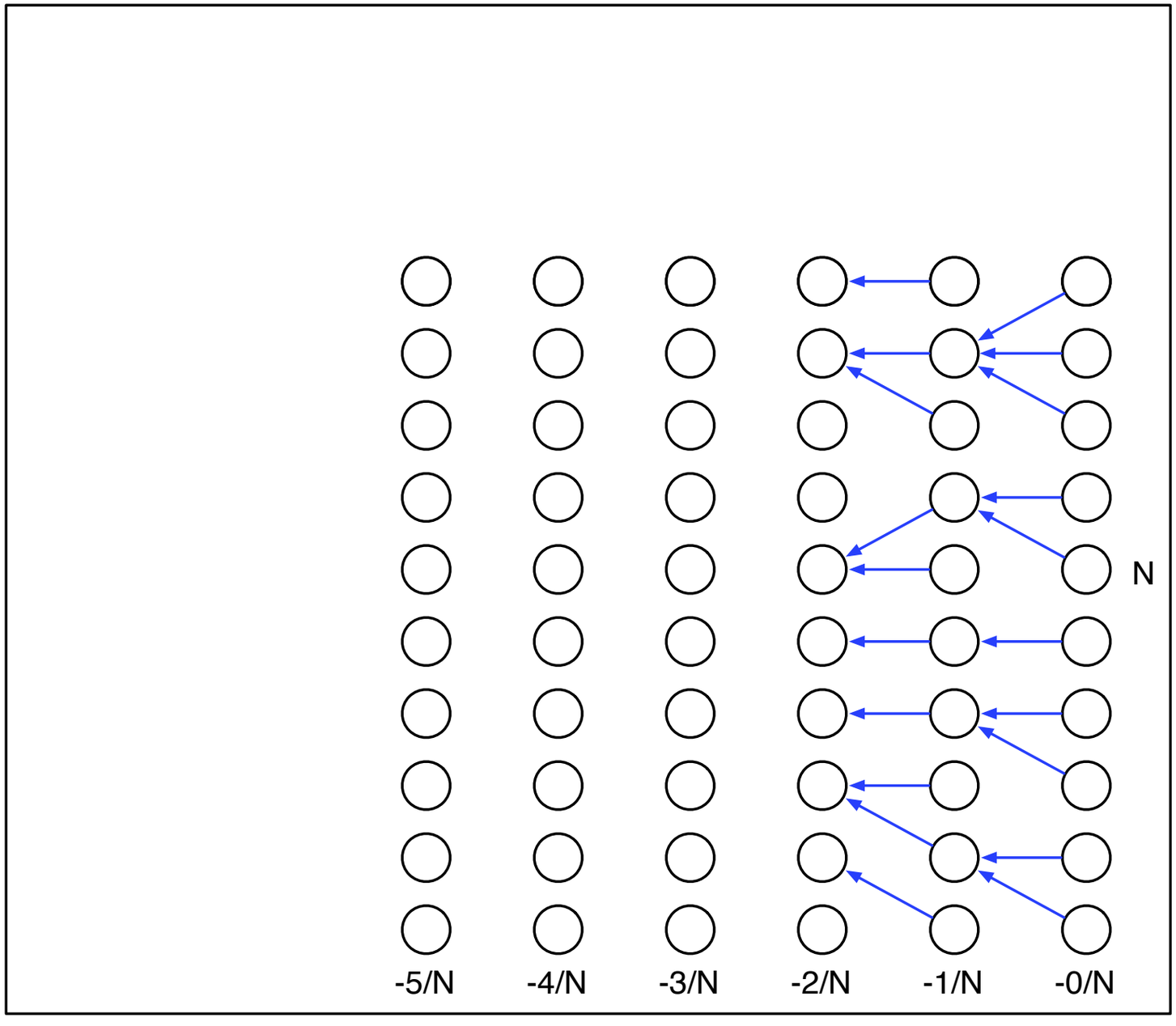
Kingman's Coalescent

- 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.



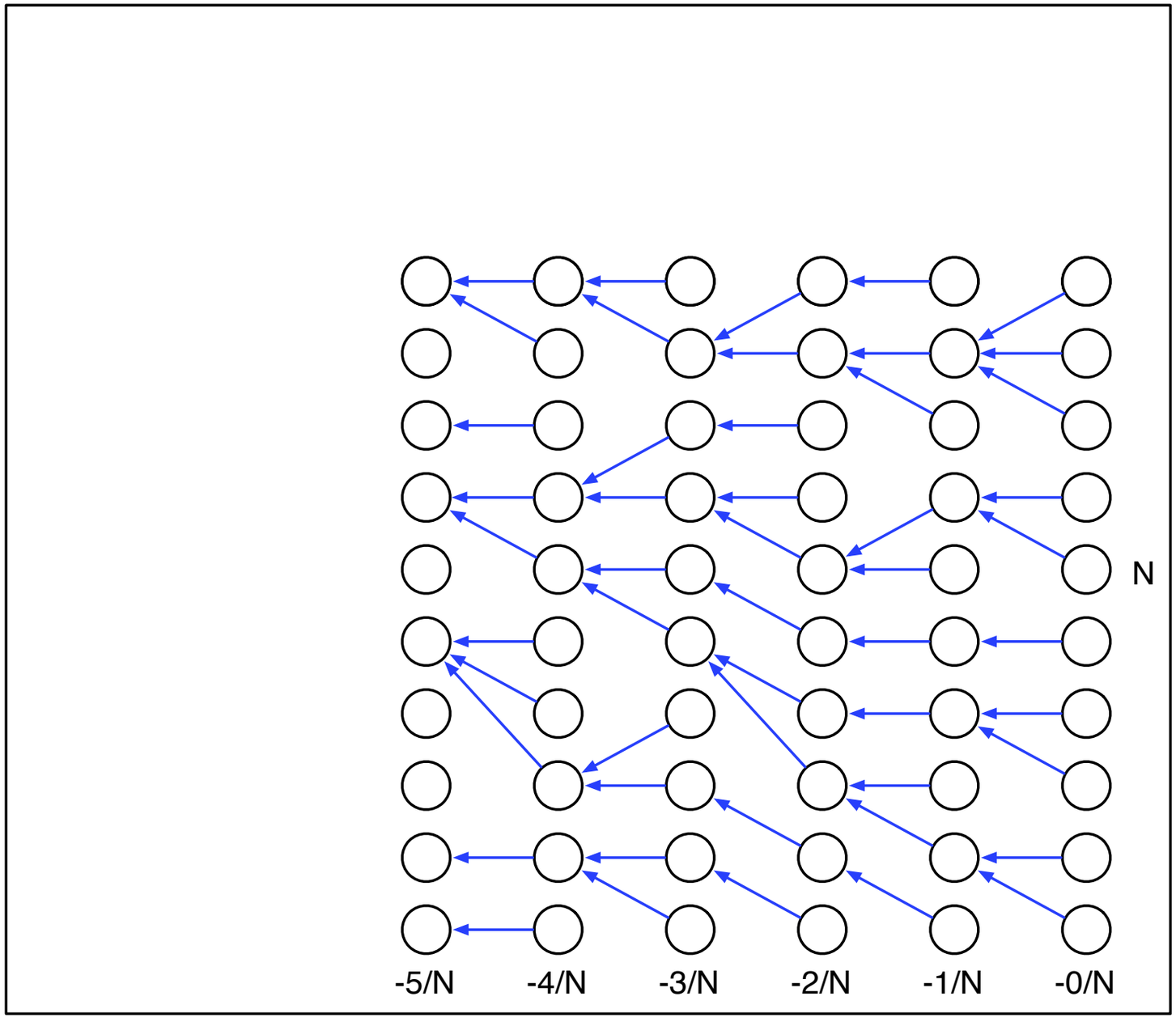
Kingman's Coalescent

- 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.



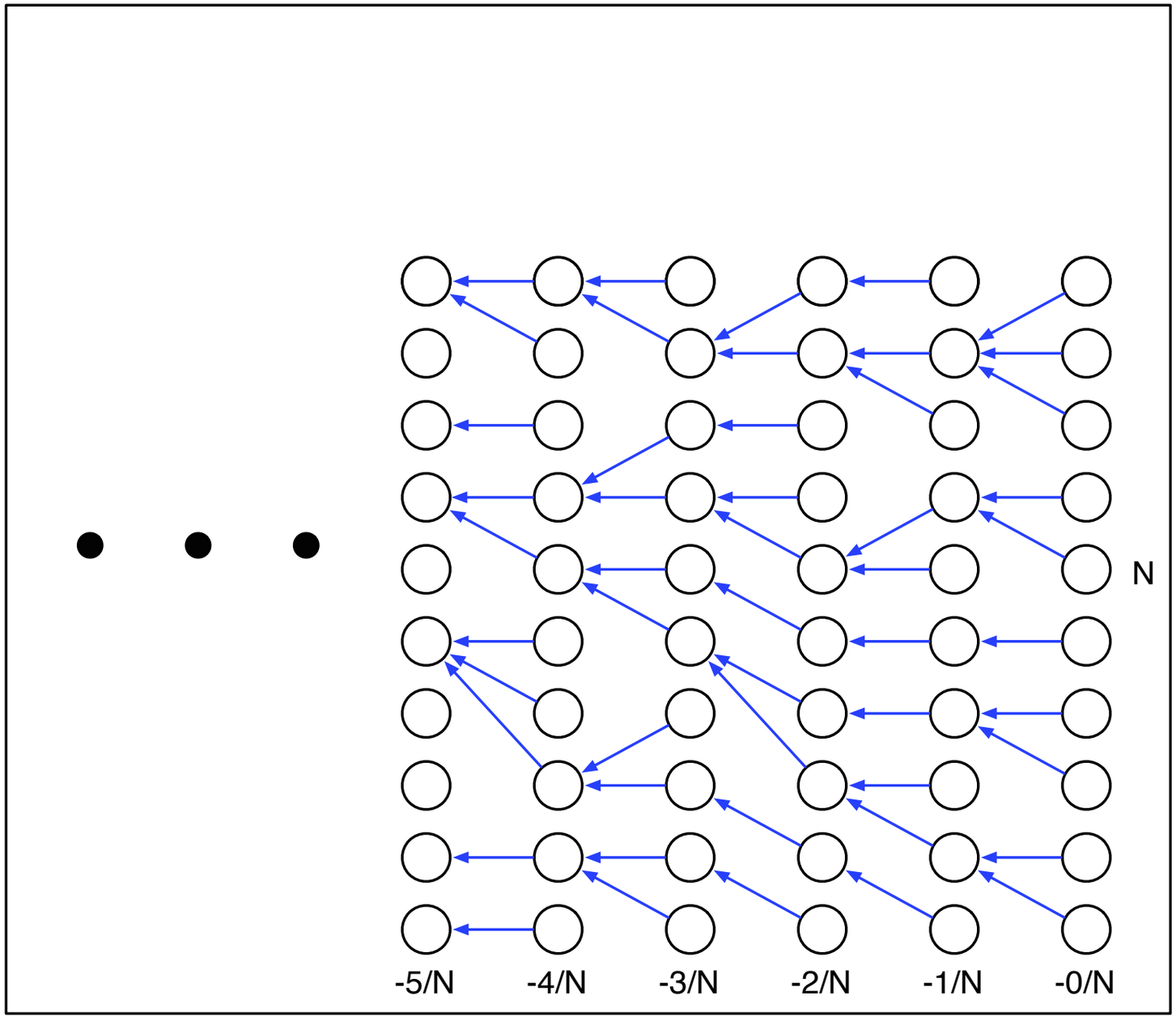
Kingman's Coalescent

- 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.



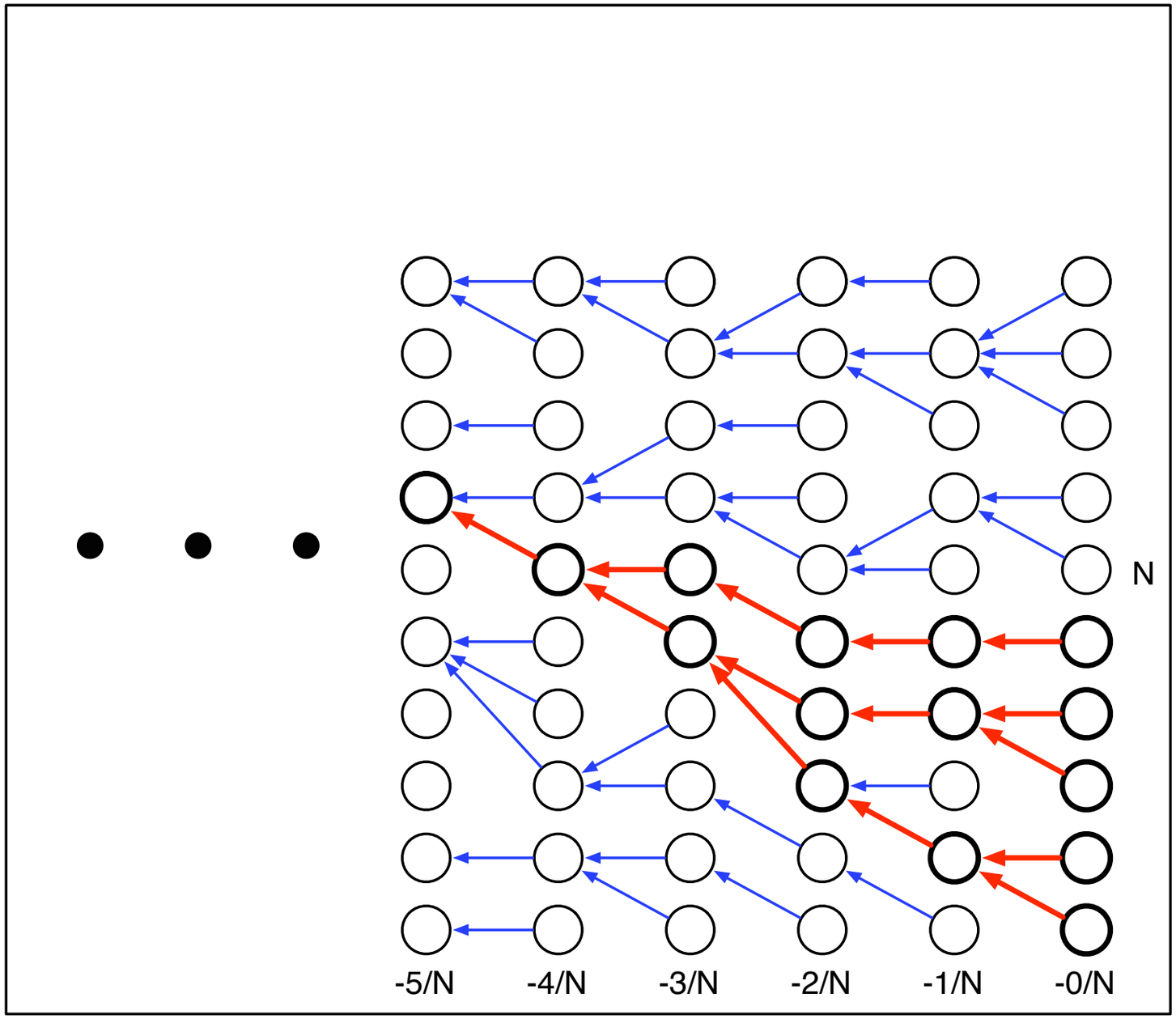
Kingman's Coalescent

- 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.



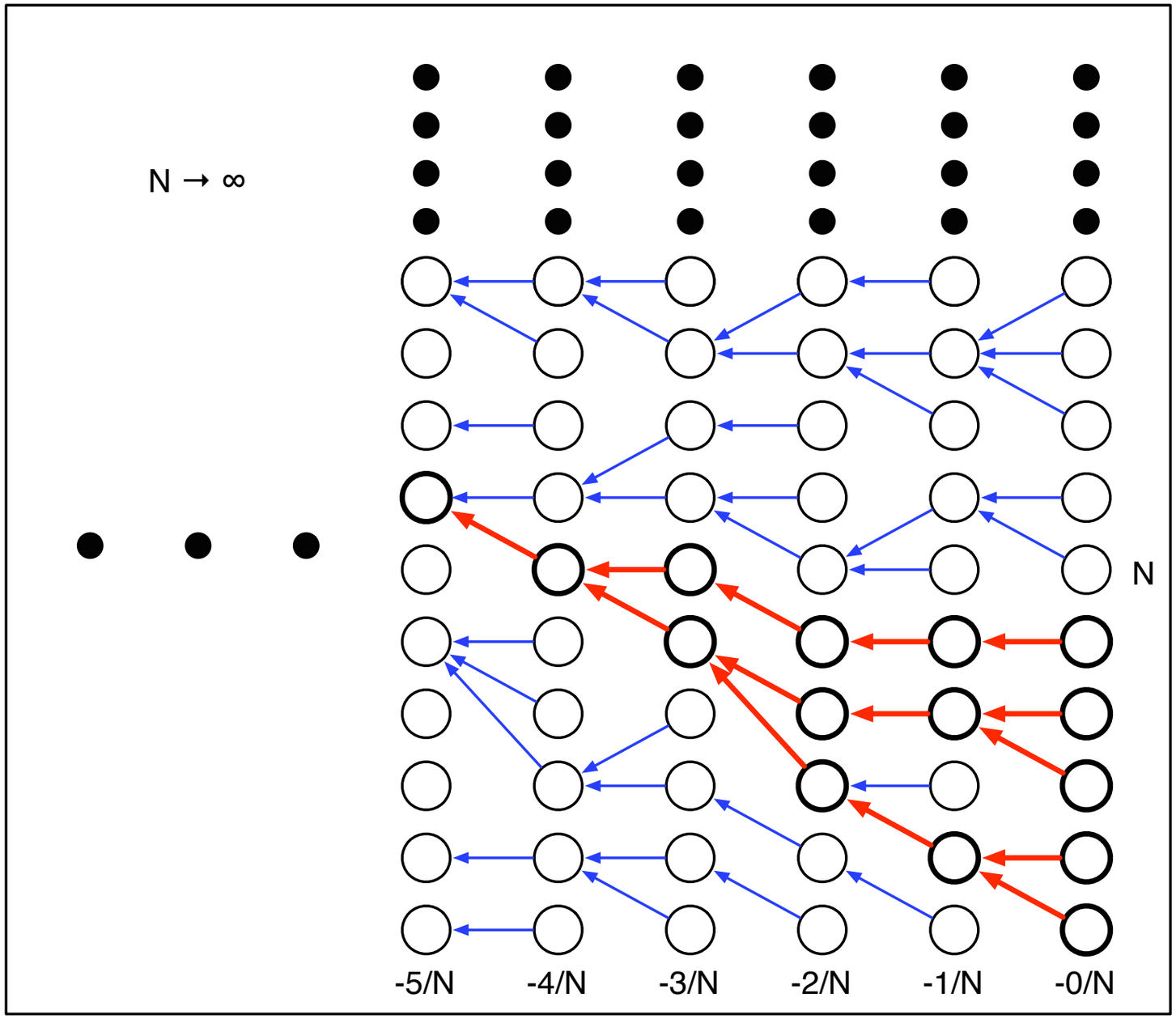
Kingman's Coalescent

- 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.

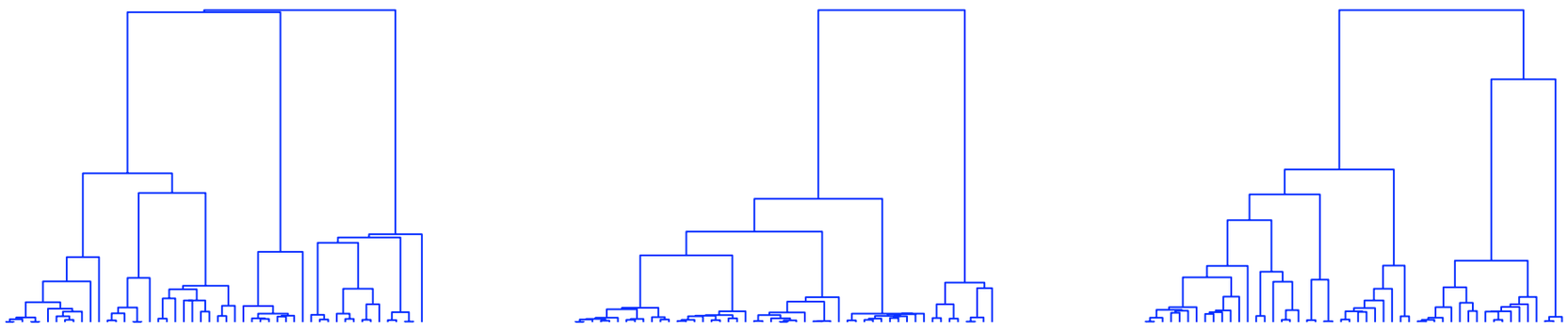


Kingman's Coalescent

- 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



Fragmentation-Coagulation Processes

[Berestycki 2004]

Overview

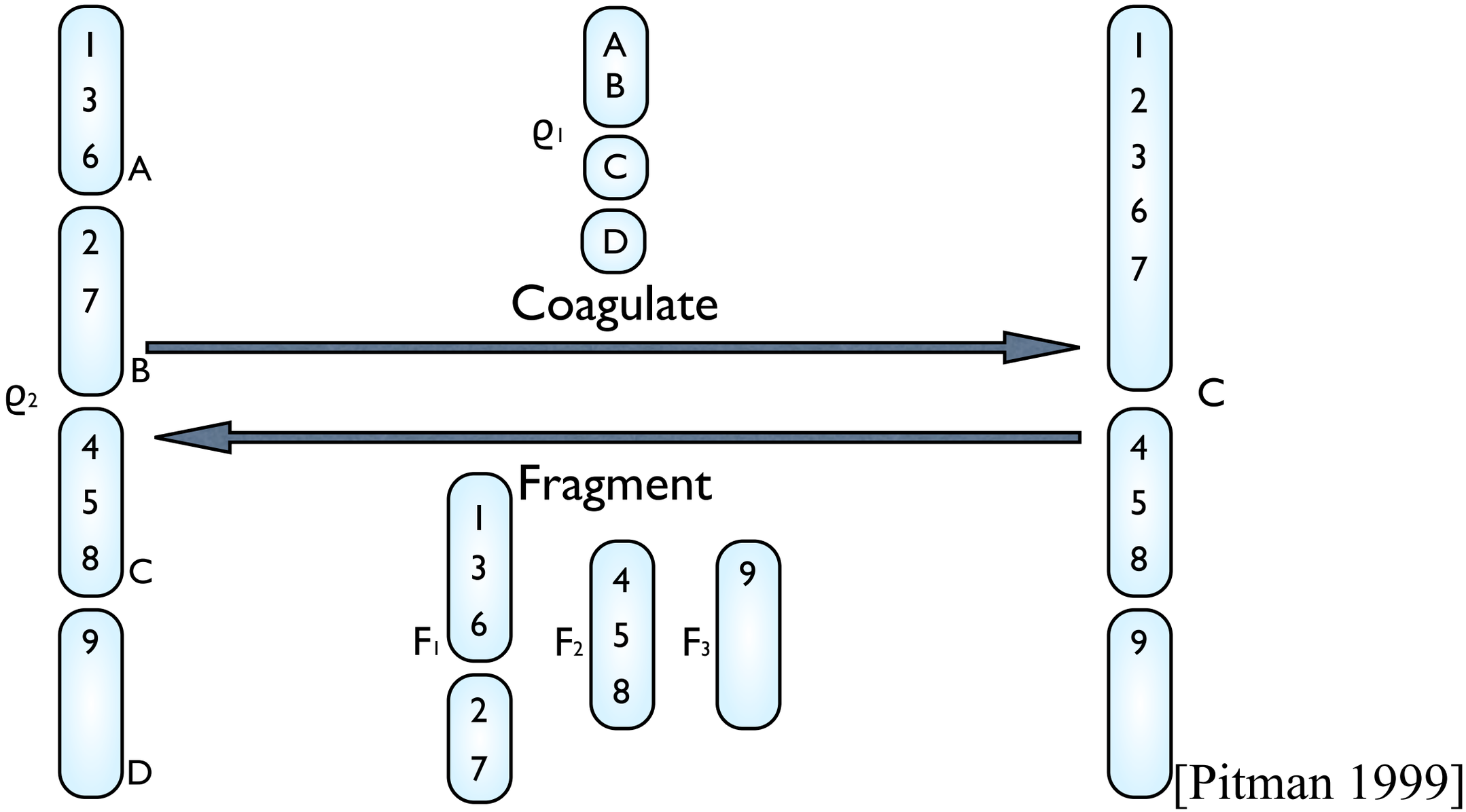
- Duality between Pitman-Yor coagulations and fragmentations.
- Using duality to construct a stationary reversible Markov chain over partitions.

Duality of Coagulation and Fragmentation

- The following statements are equivalent:

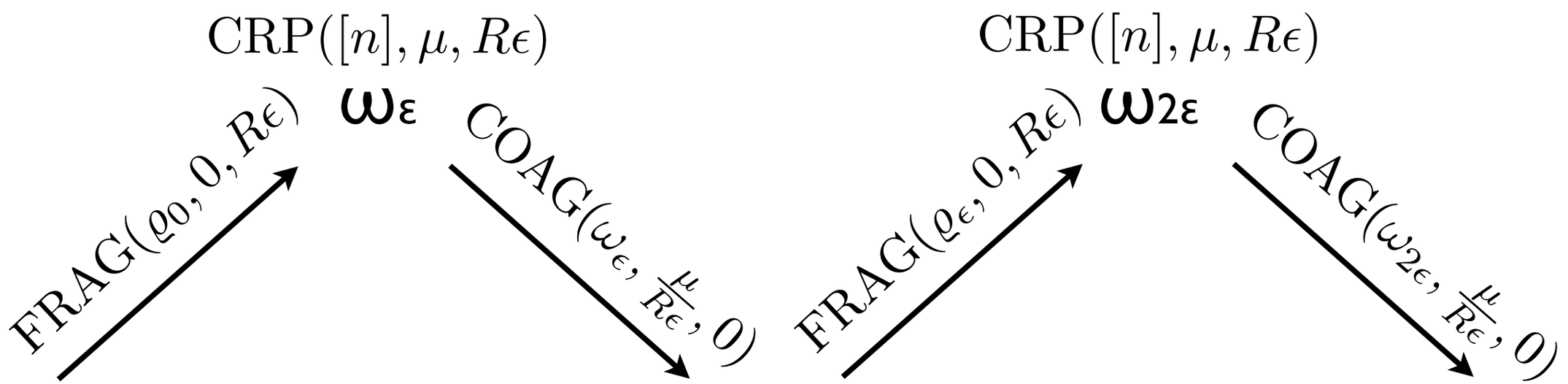
(I) $\varrho_2 \sim \mathrm{CRP}([n], d_2, \alpha d_2)$ and $\varrho_1 | \varrho_2 \sim \mathrm{CRP}(\varrho_2, d_1, \alpha)$

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



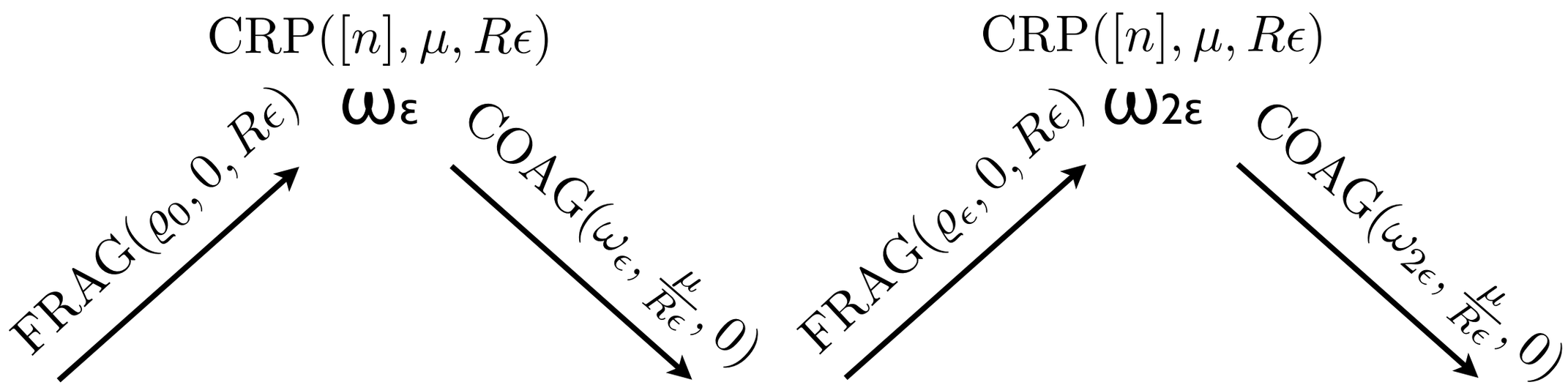
[Pitman 1999]

Markov Chain over Partitions



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

Stationary Distribution

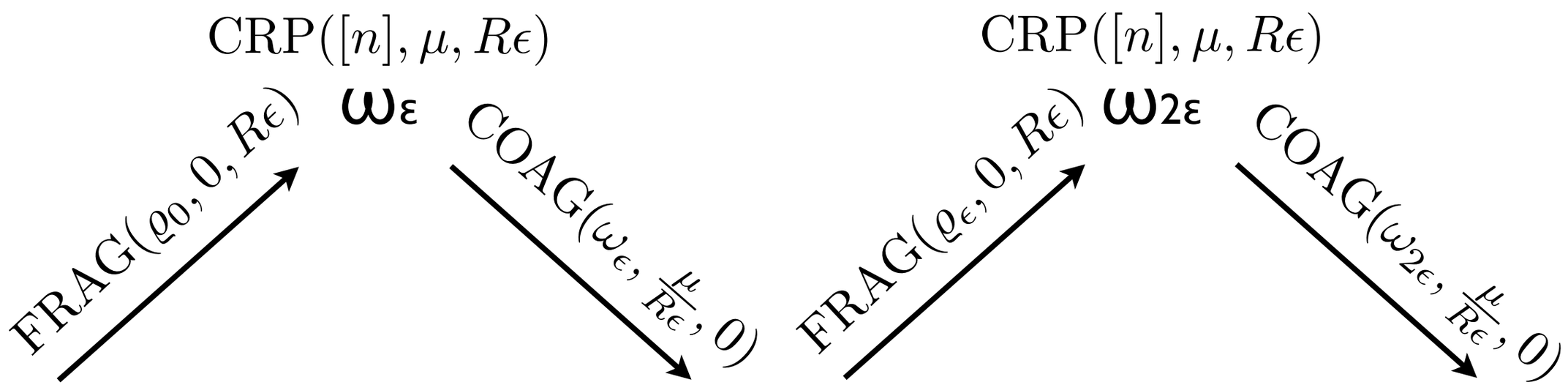


ϱ_0 $\quad$ ϱ_ε $\quad$ $\varrho_{2\varepsilon}$

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

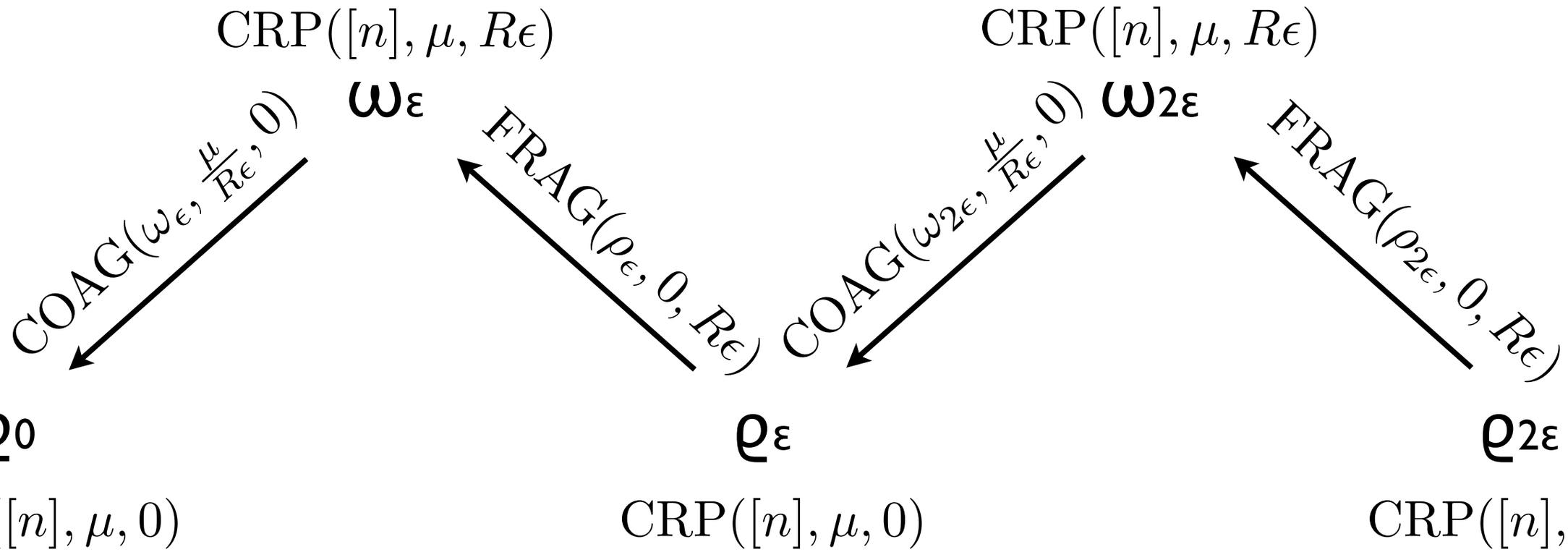
- *Stationary distribution* is a CRP with parameters μ and 0.

Exchangeability and Projectivity



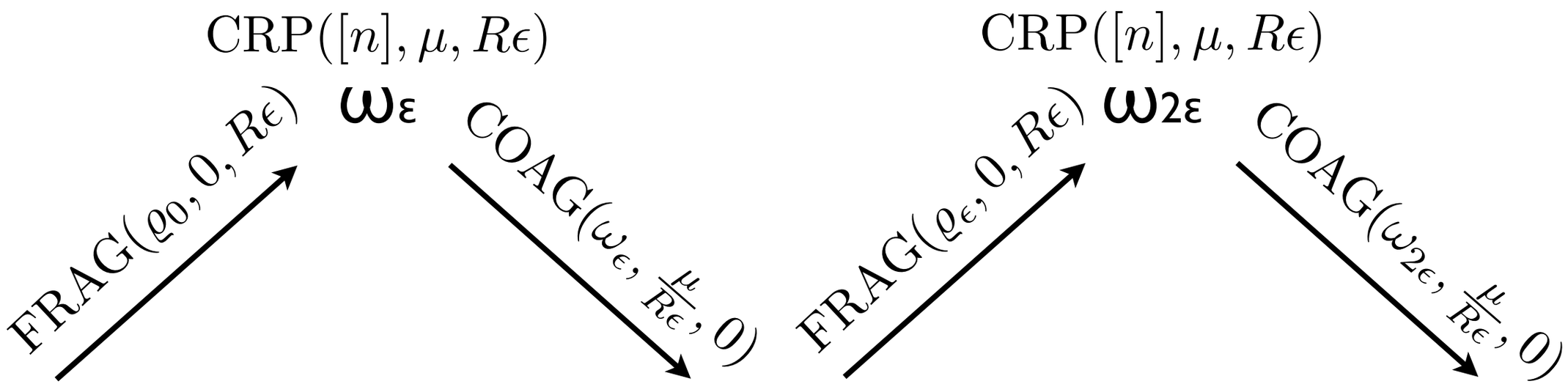
- Each π_t is exchangeable, so that the whole Markov chain is an *exchangeable process*.
- Projectivity of the Chinese restaurant process extends to the Markov chain as well.

Reversibility of Markov Chain



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

Continuum Limit

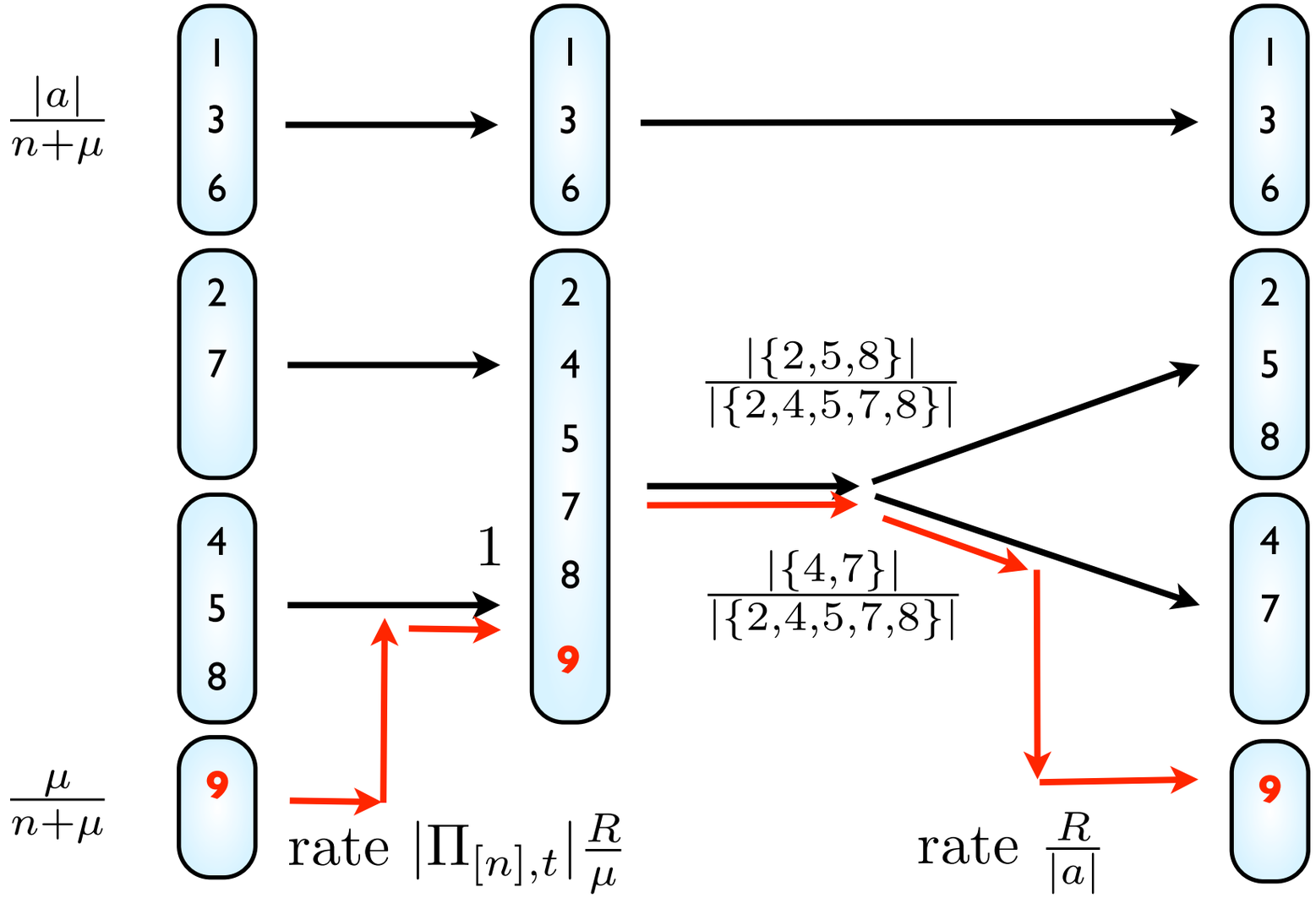


ϱ_0 $\quad\quad$ ϱ_ε $\quad\quad$ $\varrho_{2\varepsilon}$

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

- 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.

Dirichlet Diffusion Trees and Coalescents

- Rate of fragmentation is same as for Dirichlet diffusion trees with constant fragmentation rate.
- Rate of coagulation is same as for the Kingman's coalescent.
- Reversibility means that the Dirichlet diffusion tree is precisely the converse of Kingman's coalescent.

Sequence Memoizer

[Teh ACL 2006, Wood & Teh ICML 2009, Wood et al CACM 2011]

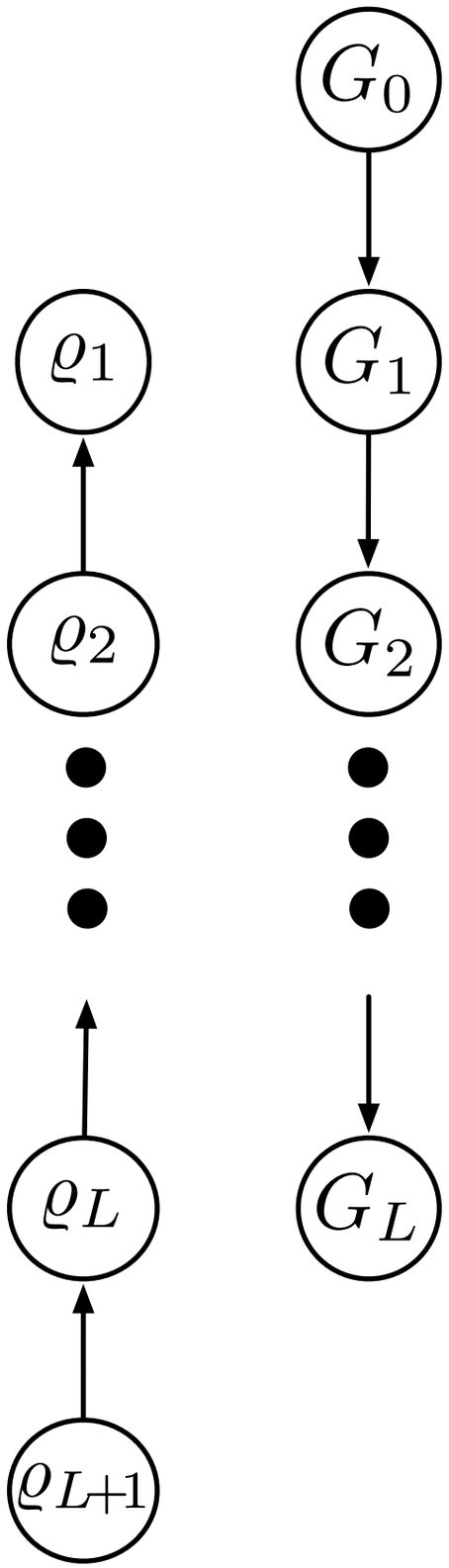
Overview

- 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.
- Generalization to ∞-gram (non-Markov) language models.
- Use of fragmentation coagulation duality to improve computational costs.

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).

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



- Generalizable to tree-structured hierarchies and **hierarchical Pitman-Yor processes**.
- 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.

[Teh et al JASA 2006, Goldwater et al NIPS 2006, Teh ACL 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:

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

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

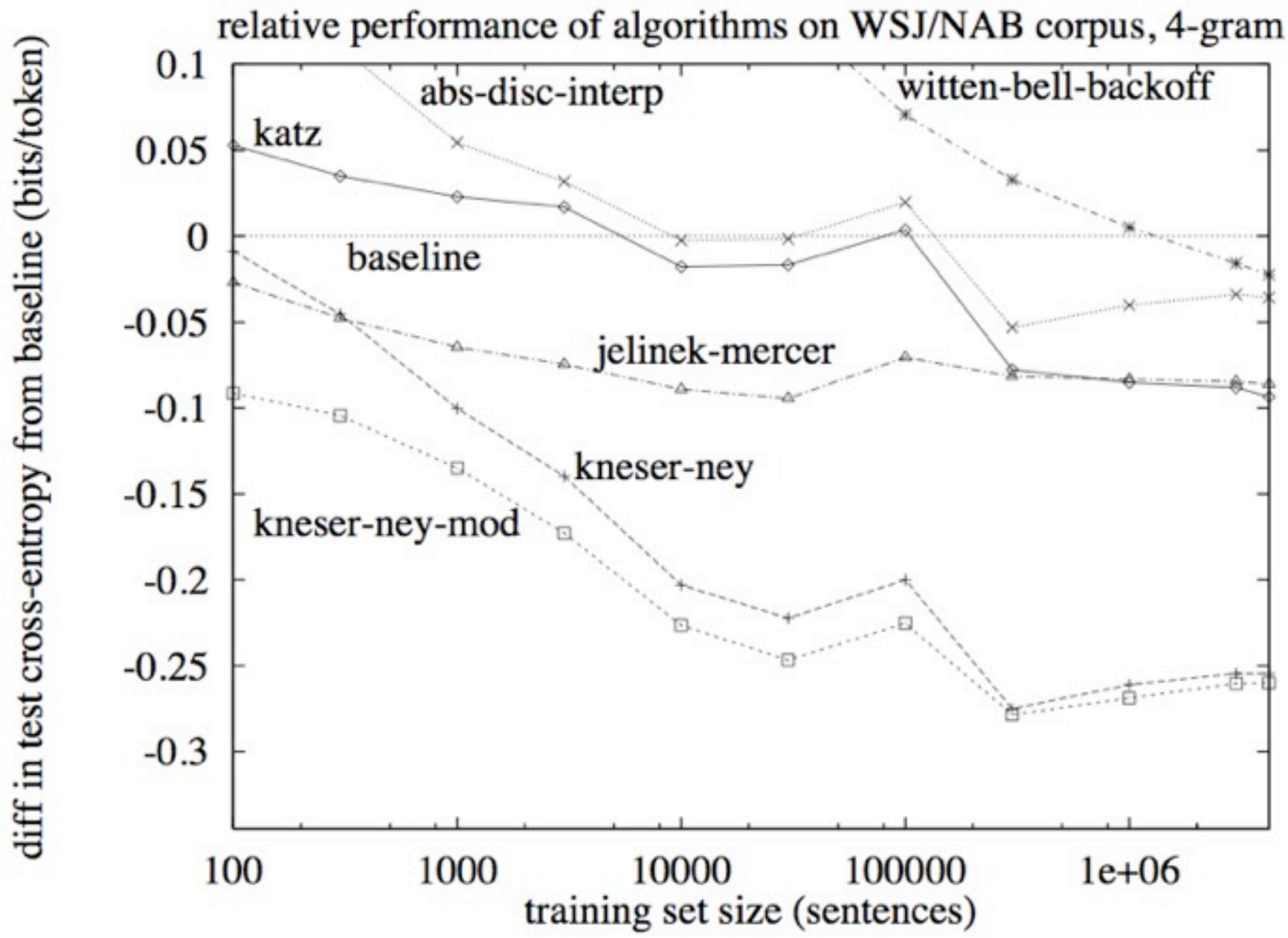
- **Smoothing** is a way of dealing with data sparsity by combining large and small models together.

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

- 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.

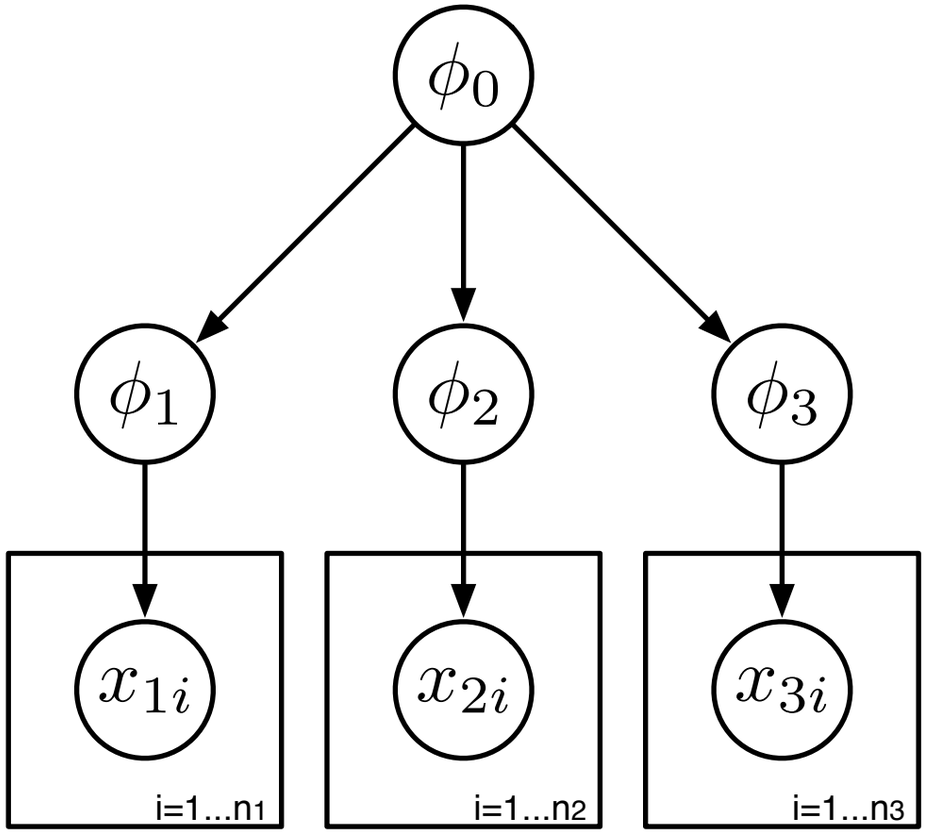
[Chen & Goodman 1999]

Hierarchical Pitman-Yor Language Models

Hierarchical Bayesian Models

- **Hierarchical Bayesian modelling** an important overarching theme in modern statistics [Gelman et al, 1995, James & Stein 1961].

- In machine learning, have been used for multitask learning, transfer learning, learning-to-learn and domain adaptation.



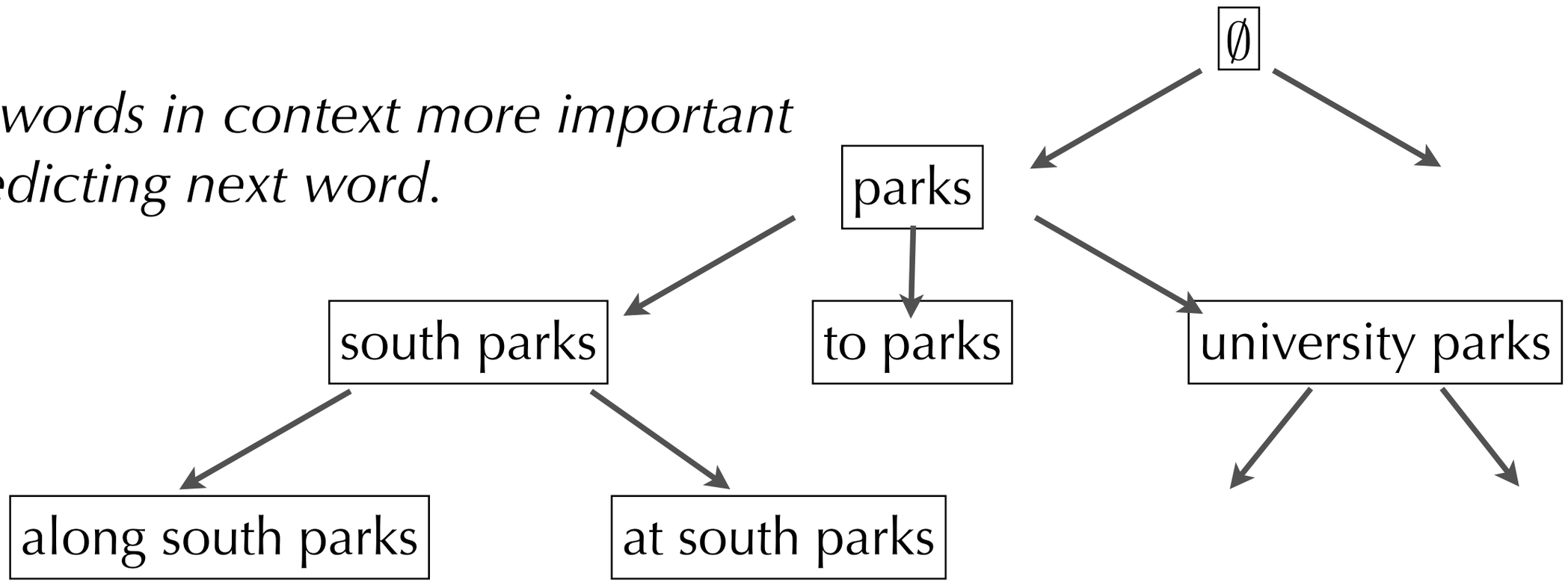
Context Tree

- **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

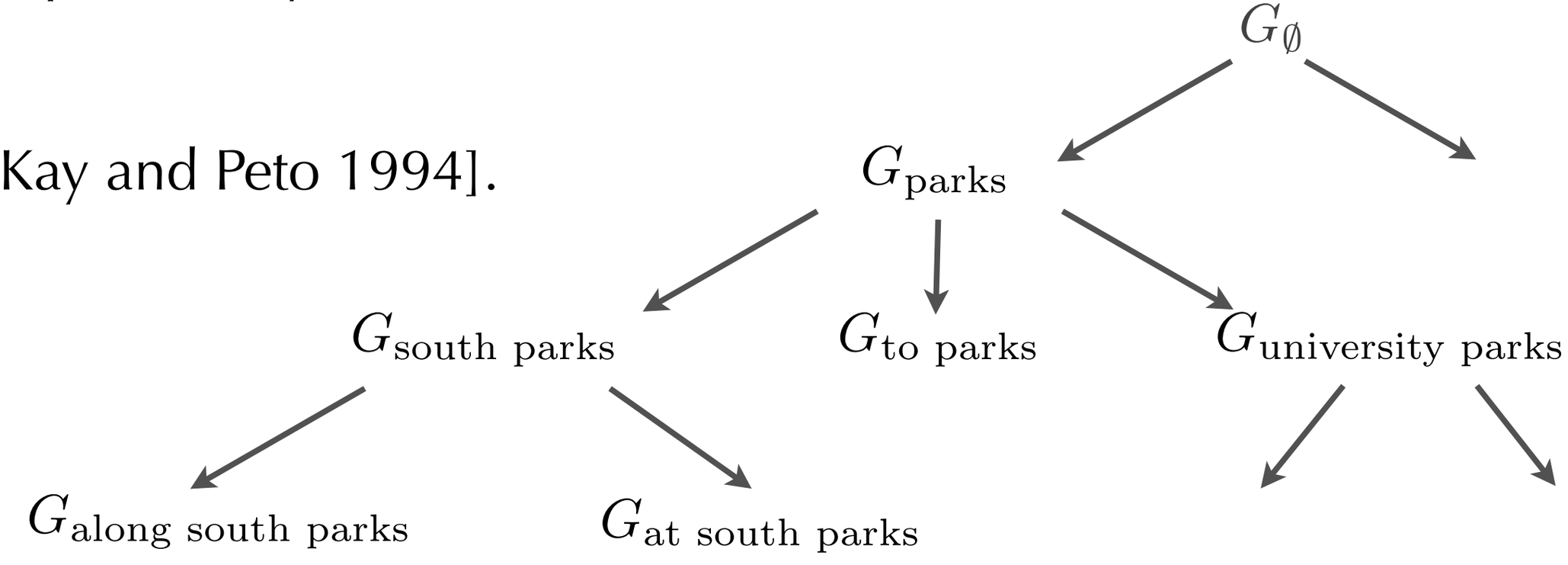
- Parametrize the conditional probabilities of Markov model:

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

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

- G_u is a probability vector associated with context u.

- [MacKay and Peto 1994].



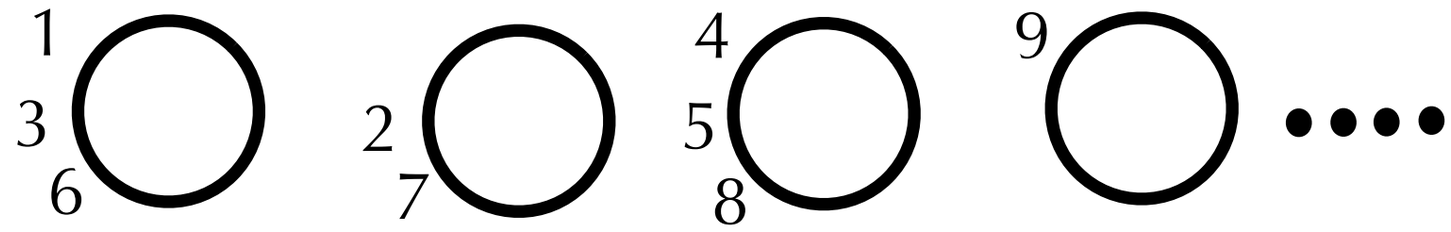
Hierarchical Dirichlet Language Models

- What is $P(G_u|G_{\mathrm{pa}(u)})$? [MacKay and Peto 1994] proposed using 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 [Perman, Pitman and Yor 1992], [Pitman and Yor 1997], [Ishwaran and James 2001].

Two-parameter Chinese Restaurant Processes



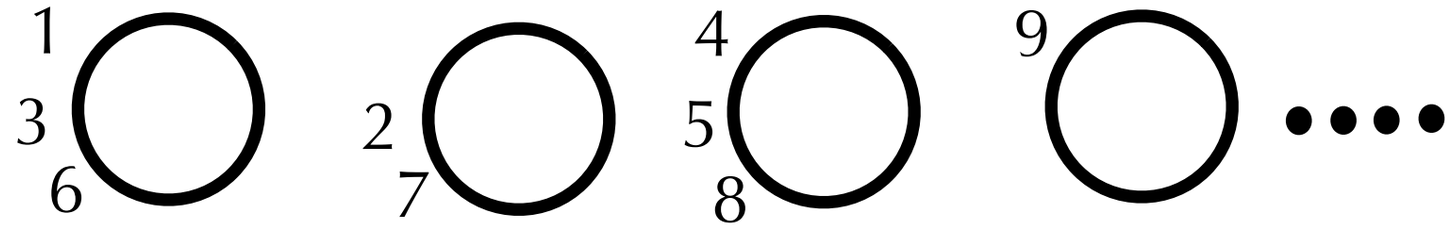
$$P(\text{sit at table } c) = \frac{n_c - d}{\alpha + \sum_{c \in \varrho} n_c} \qquad P(\text{sit at new table}) = \frac{\alpha + d|\varrho|}{\alpha + \sum_{c \in \varrho} n_c}$$

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

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

[Perman et al 1992, Pitman & Yor AoP 1997]

Two-parameter Chinese Restaurant Processes



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

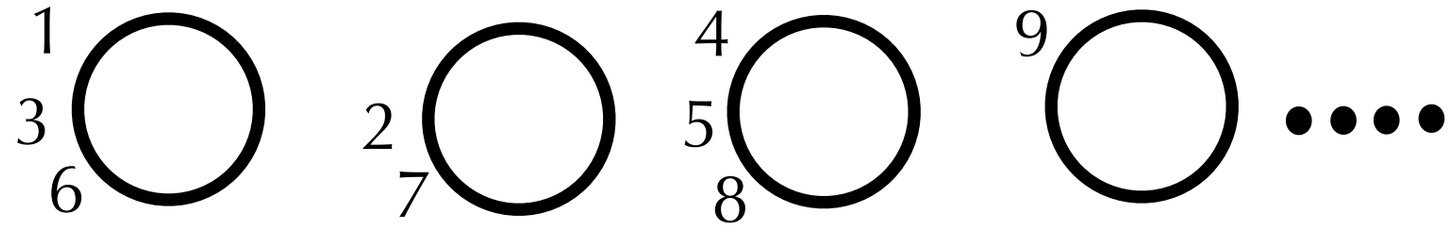
- The **two-parameter Chinese restaurant process** CRP([n],d,α) is a distribution over $\mathcal{P}_{[n]}$: ($0 \leq d < 1$, $\alpha > -d$)

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

- These are also projective and exchangeable distributions.

[Perman et al 1992, Pitman & Yor AoP 1997]

Two-parameter Chinese Restaurant Processes



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

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

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

- These are also projective and exchangeable distributions.
- De Finetti measure is the **Pitman-Yor process**, which is a generalization of the Dirichlet process.

[Perman et al 1992, Pitman & Yor AoP 1997]

Power Law Properties

- Chinese restaurant process:

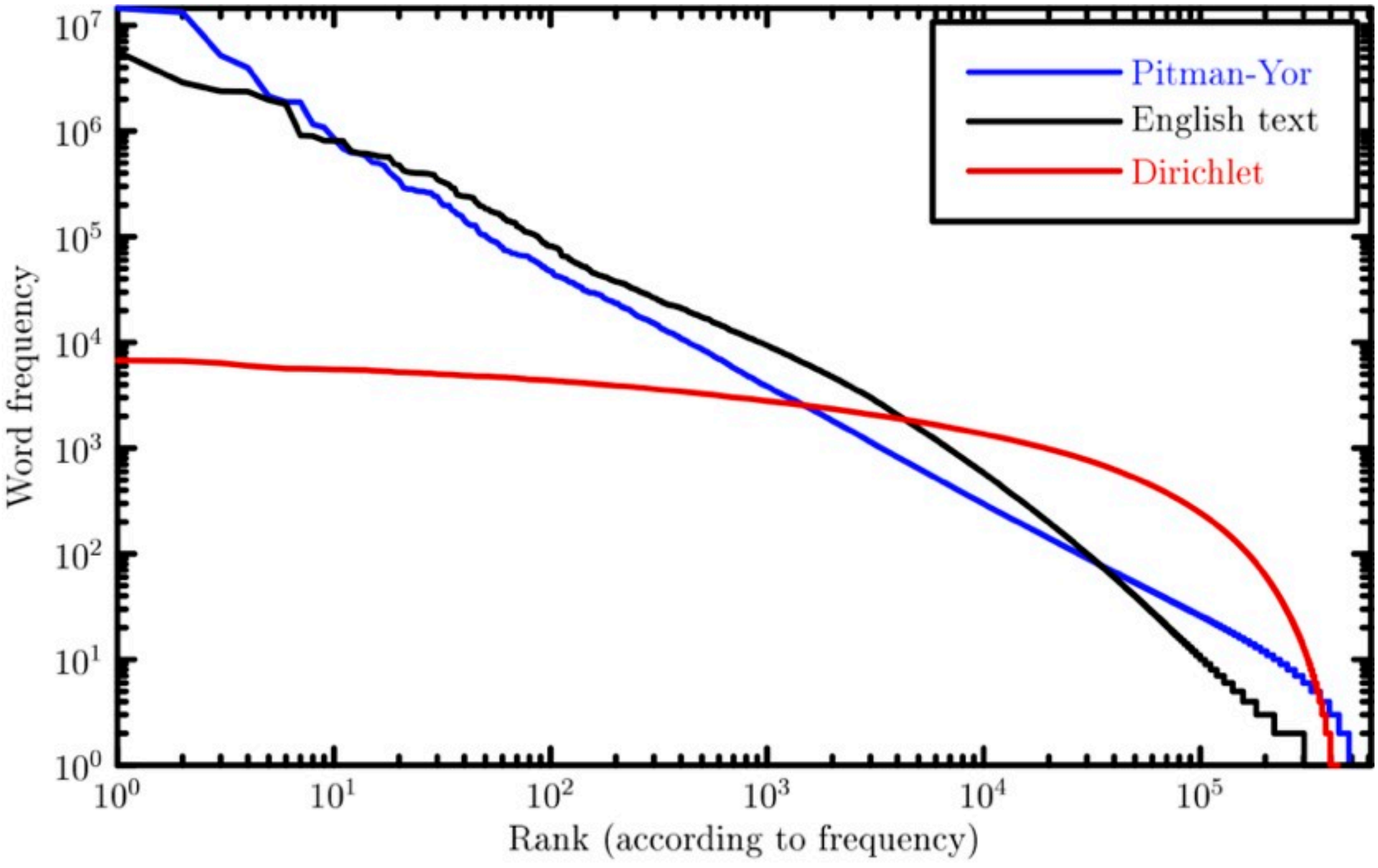
$$p(\text{sit at table } c) \propto n_c - d$$
$$p(\text{sit at new table}) \propto \alpha + d|\rho|$$

  - Small number of large clusters;
  - Large number of small clusters.
  - Customers = word instances, tables = word types.

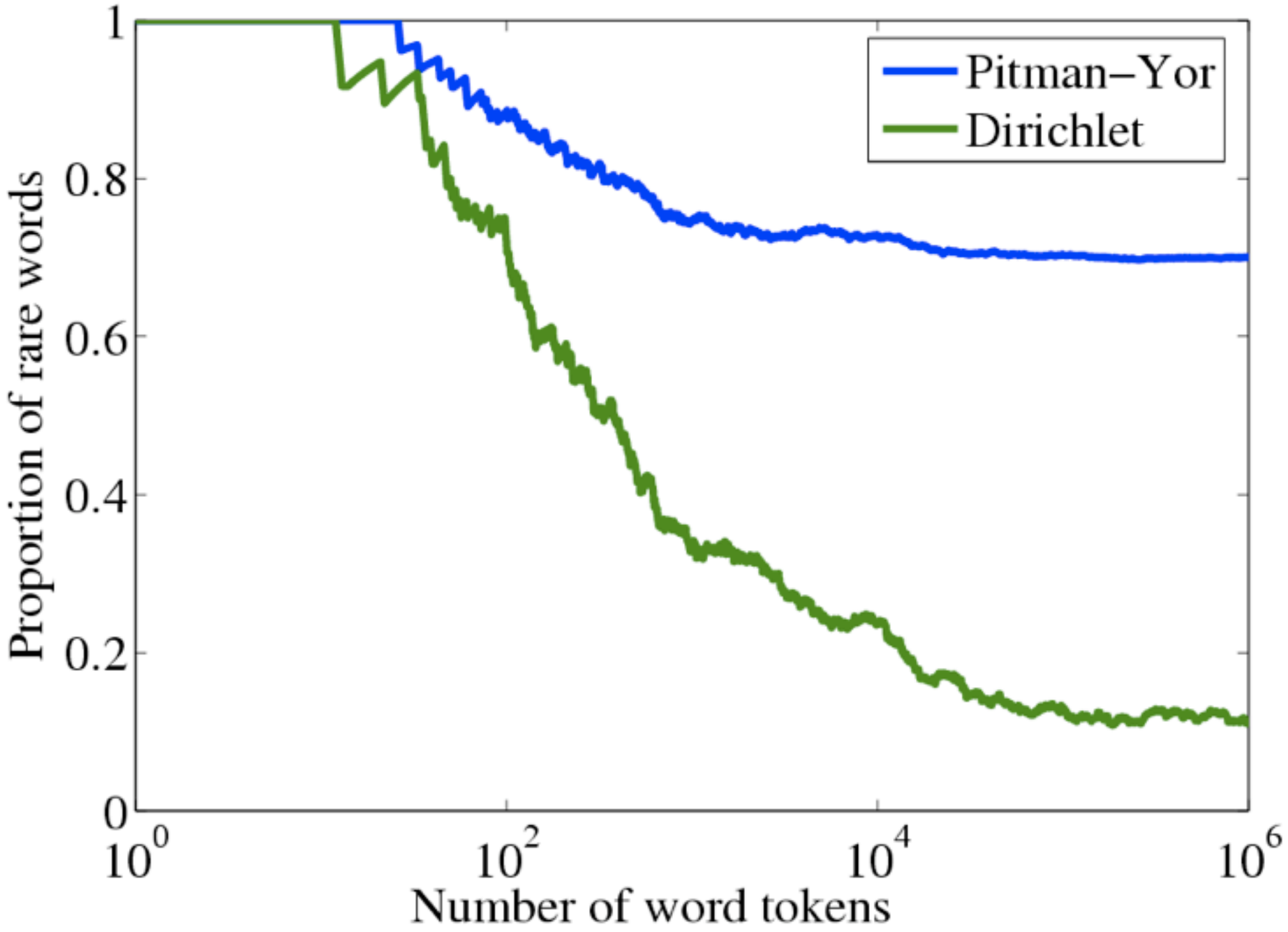
- This is more suitable for languages than Dirichlet distributions.

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

Power Law Properties



Power Law Properties



Hierarchical Pitman-Yor Language Models

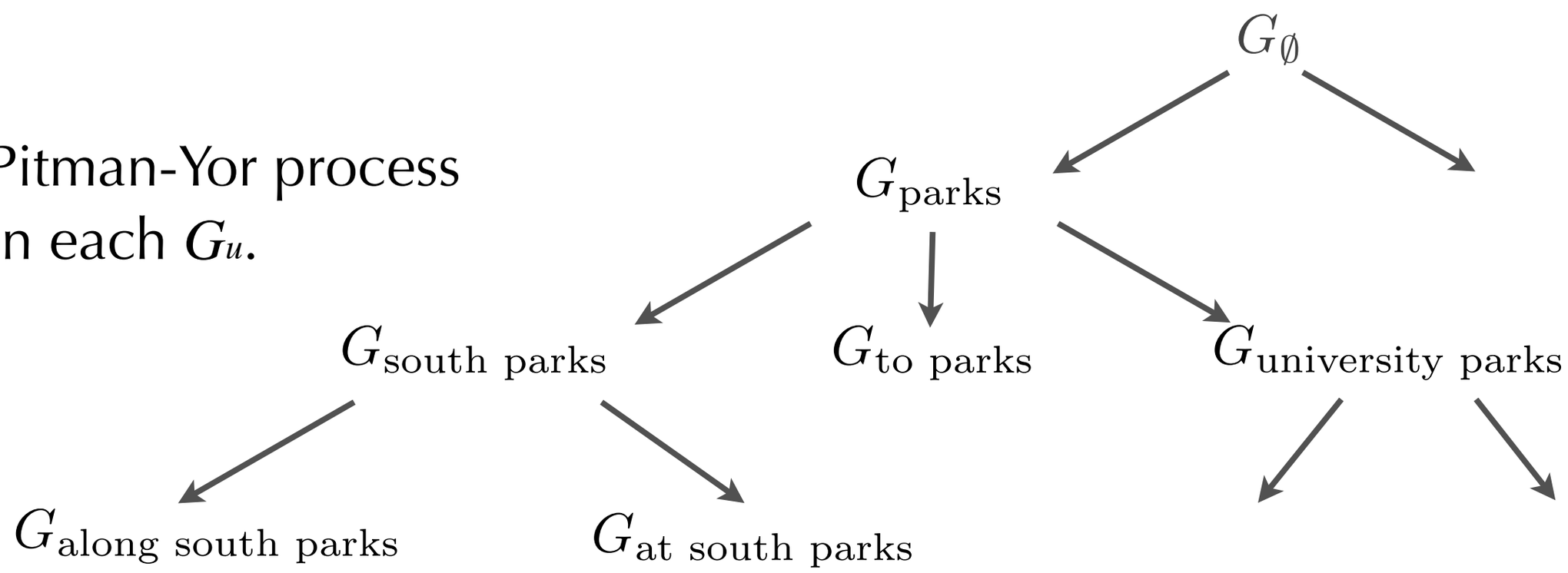
- Parametrize the conditional probabilities of Markov model:

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

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

- G_u is a probability vector associated with context u.

- Place Pitman-Yor process prior on each G_u.



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.

Non-Markov Language Models

Markov Models for Language and Text

- Usually makes a Markov assumption to simplify model:

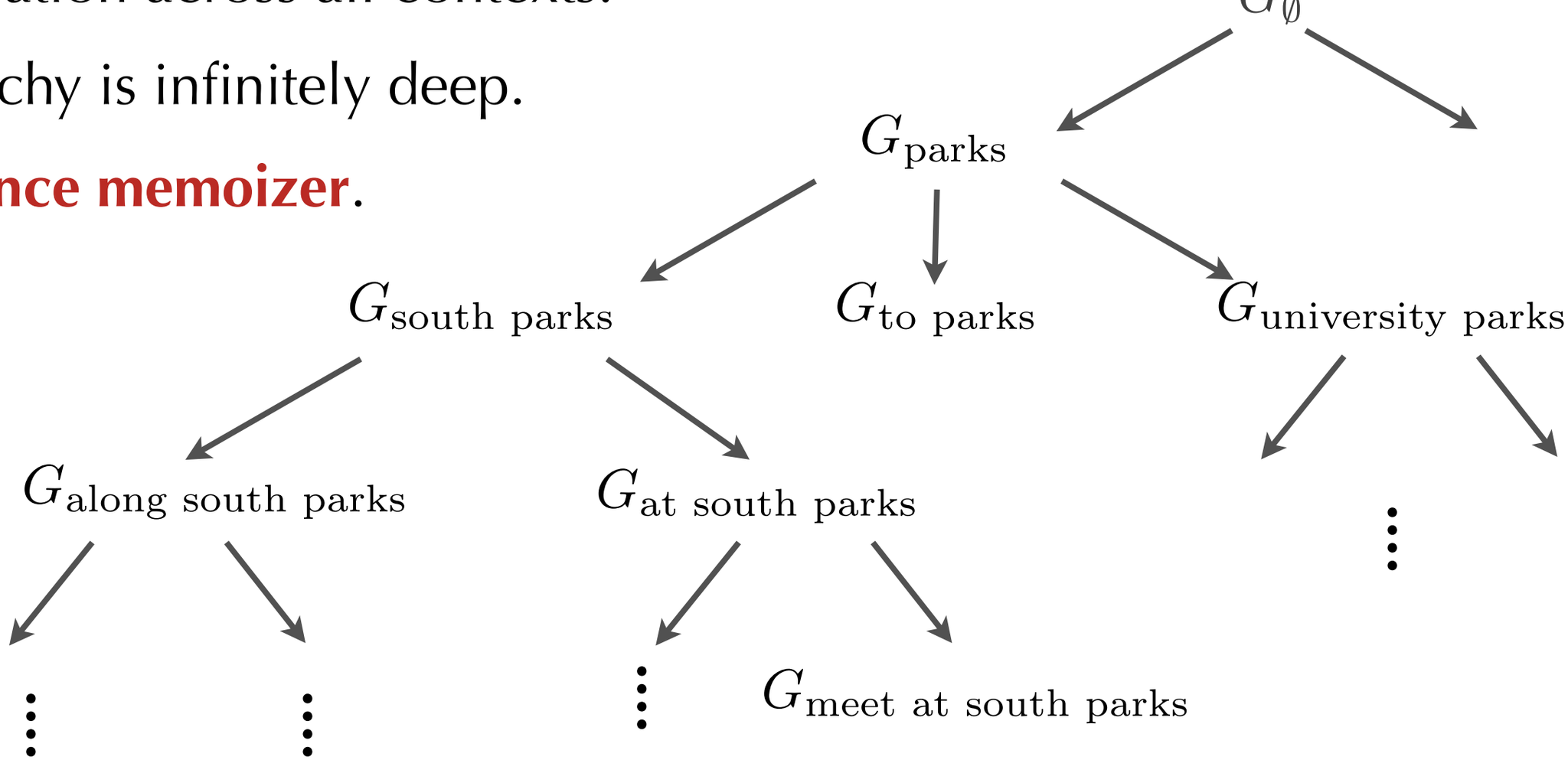
P(south parks road) ~
P(south)*
P(parks | south)*
P(road | south parks)

- 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

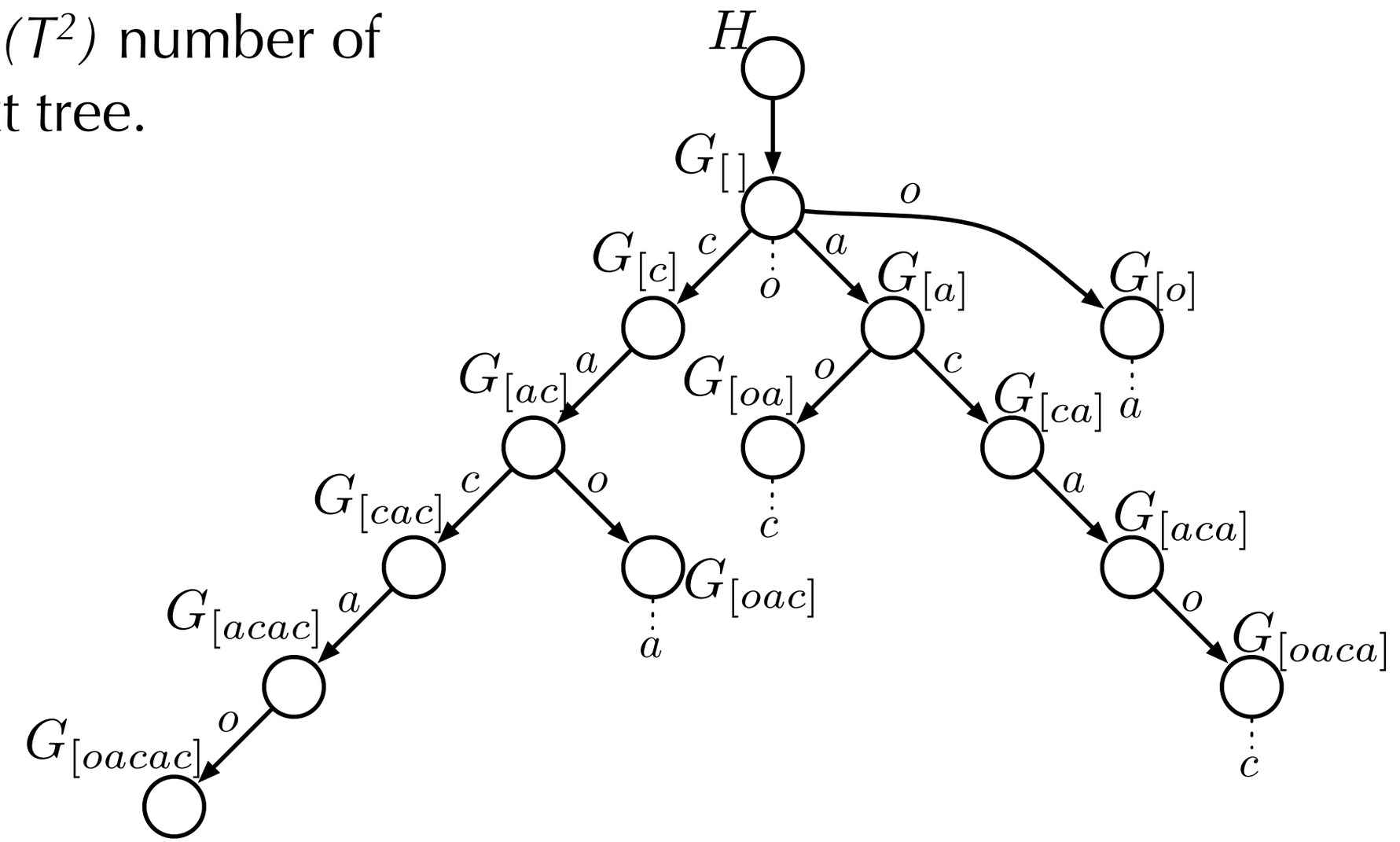
- Model the conditional probabilities of each possible word occurring after each possible context (of unbounded length).

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



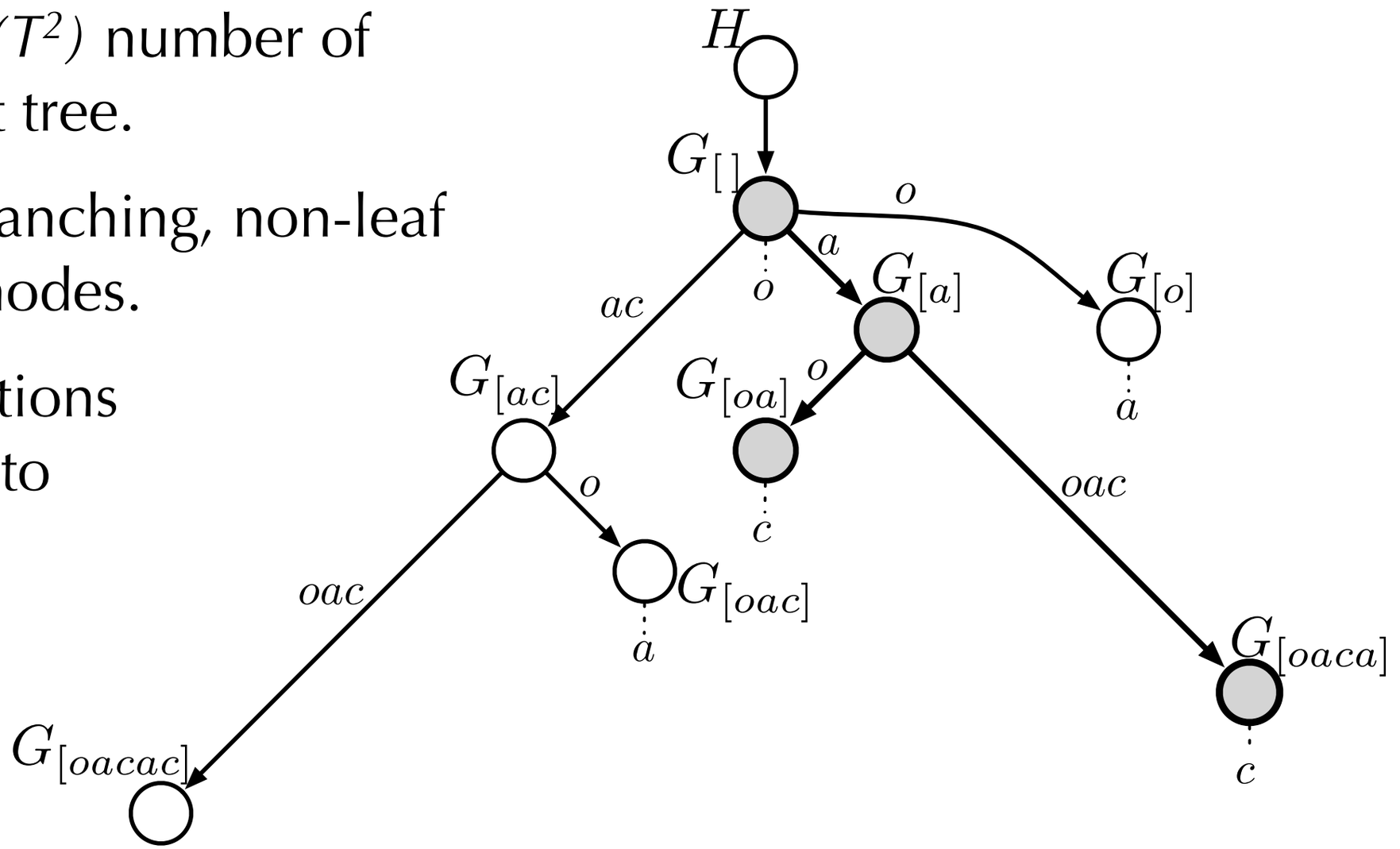
Model Size: Infinite -> $O(T^2)$

- 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 -> $O(T^2)$ -> 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 still Pitman-Yor due to closure property.

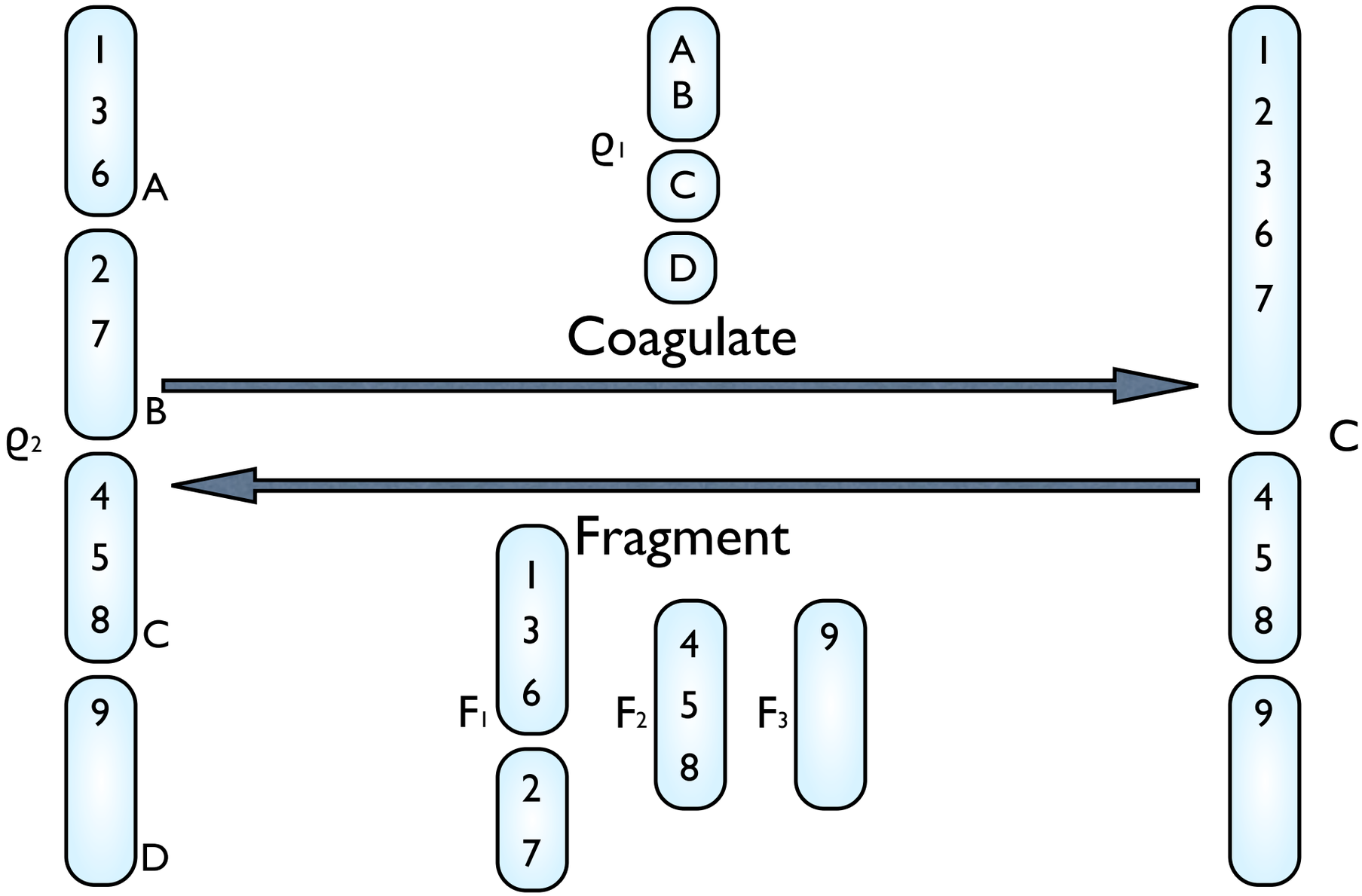


Duality of Coagulation and Fragmentation

- The following statements are equivalent:

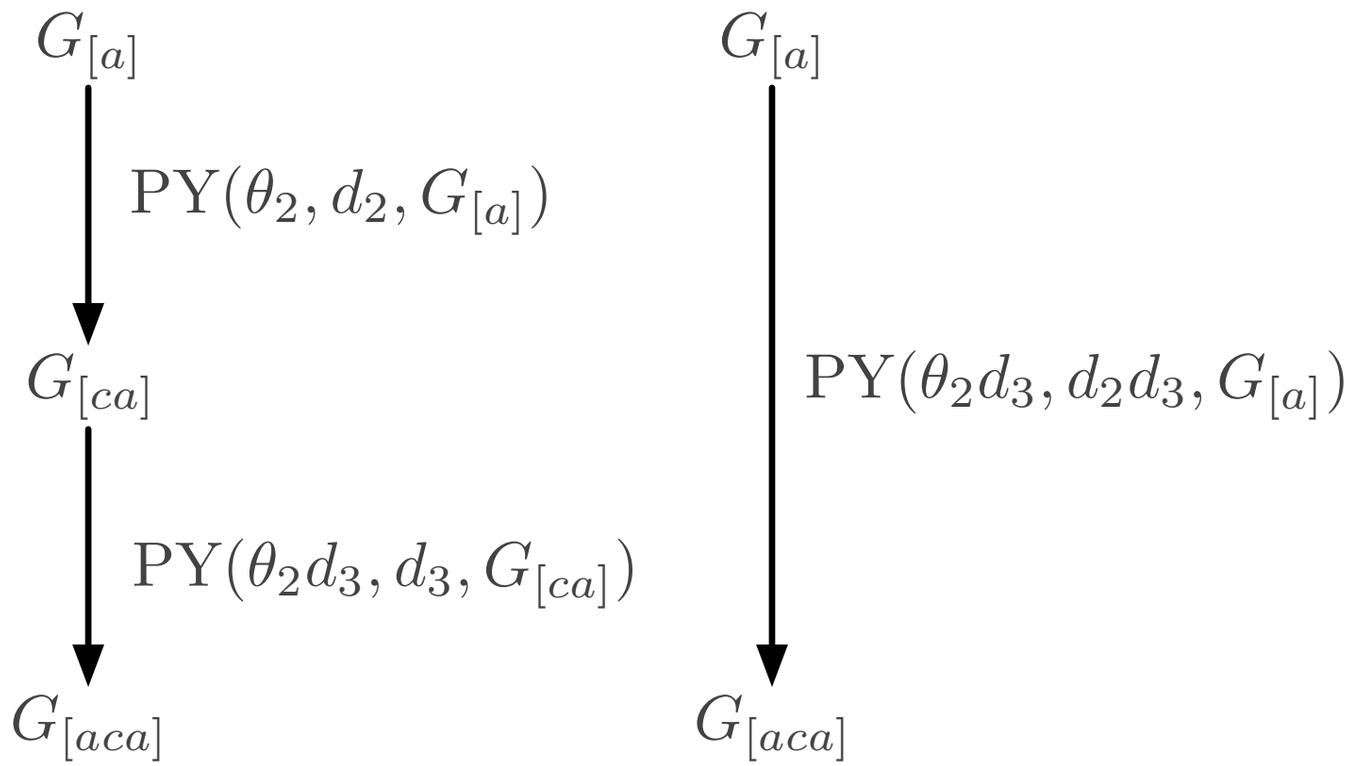
(I) $\varrho_2 \sim \mathrm{CRP}([n], d_2, \alpha d_2)$ and $\varrho_1 | \varrho_2 \sim \mathrm{CRP}(\varrho_2, d_1, \alpha)$

(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

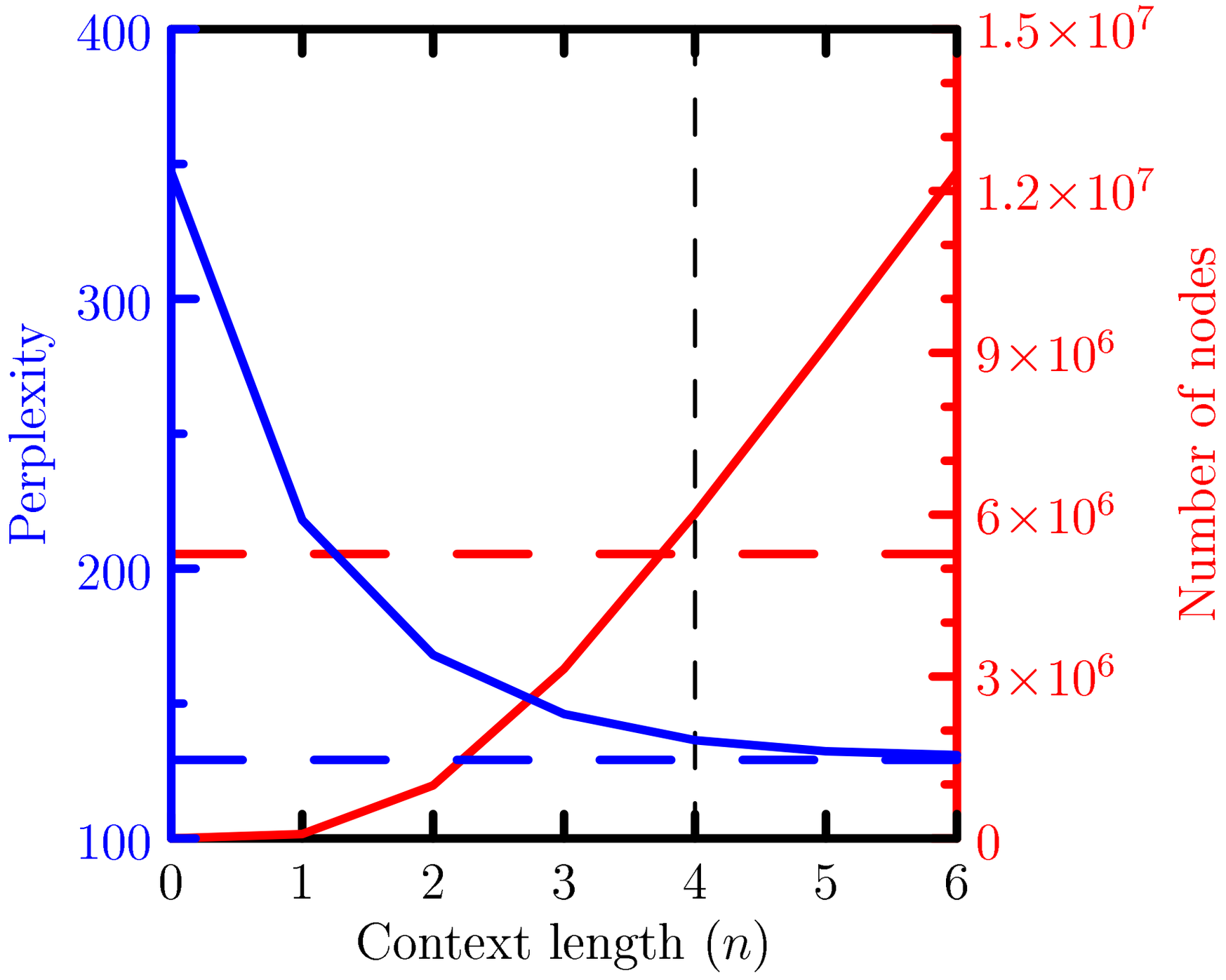
- Marginalizing out internal Pitman-Yor processes is equivalent to coagulating the corresponding Chinese restaurant processes.



- 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 $G_{[aca]}$.

[Gasthaus & Teh NIPS 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.

A Few Final Words

Summary

- Introduction to Bayesian learning and Bayesian nonparametrics.

- Chinese restaurant processes.
- Fragmentations and coagulations.

- Unifying view of random tree models as Markov chains of fragmenting and coagulating partitions.
- Fragmentation-coagulation processes.

- Hierarchical Dirichlet and Pitman-Yor processes, sequence memoizer.

What Were Not Covered Here

- Bayesian nonparametrics in computational linguistics (Mark Johnson).
- Gaussian processes, Indian buffet processes (Zoubin Ghahramani).
- Nonparametric Hidden Markov Models [see Ghahramani CoNLL 2010].
- Dependent and hierarchical processes [see Dunson BNP 2010, Teh & Jordan BNP 2010].

- Foundational issues, convergence and asymptotics.
- Combinatorial stochastic processes and their relationship to data structures and programming languages.
- Relational models, topic models etc.

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.