Binary to Bushy: Bayesian Hierarchical Clustering with the Beta Coalescent

Yuening Hu[1], **Jordan Boyd-Graber**[2], **Hal Daumè III**[1], **Z. Irene Ying**[3]
[1]Computer Science, [2]iSchool and UMIACS, [3]Agricultural Research Service
[1,2]University of Maryland, [3]Department of Agriculture
ynhu@cs.umd.edu, {jbg,hal}@umiacs.umd.edu, irene.ying@gmail.com

This document provides additional information for the submission: a reference for notation, the Dirichlet process mixture models sampling (DPMM) with a Gaussian base distribution, and the procedure for generating synthetic data (including examples of generated and reconstructed trees).

1 Review of Notation

n	the number of observations
m	the total number of coalescent events
t_i	the time when the i^{th} coalescent event happens
n_i	the number of nodes at time t_i
δ_i	the time duration between t_i and t_{i-1}, and $t_i = t_{i-1} - \delta_i$
π	a tree structure
λ_n^k	the rate at which k out of n nodes merge into a parent node
λ_n	the total rate of any children set merging of n nodes
γ	a fraction of nodes coalescing
$\Lambda(d\gamma)$	a finite measure on $[0, 1]$
α	the parameter of beta distribution
ρ_i	a node
$\rho_{\vec{c}_i}$	a set of children nodes of node ρ_i
$\lvert\rho_{\vec{c}_i}\rvert$	the size of children set $\rho_{\vec{c}_i}$
$\mathbf{x}$	the data observations
y_i	the feature vector associated with node ρ_i
$p_0(y_i)$	the initial distribution of node ρ_i with feature y_i
$\kappa_{t_i t_b}(y_i, y_b)$	the transition kernel from node feature y_i formed at t_i to node feature y_b formed at t_b
μ	the mutation rate
$M_{\rho_i}(y_i)$	the message of node ρ_i with node feature y_i
Z_{ρ_i}	the local normalizer at node ρ_i
$Z_{-\infty}$	the normalizer at $-\infty$
θ_i	the subtree structure of all observations at time t_i
s	the particle index
w_i^s	the weight of particle s at time t_i
f	the proposal distribution
Z_0	the normalizer of local normalizers Z_{ρ_i}
Ω_i	a restricted set of children sets at time t_i
ω_{ij}	the j^{th} children set of Ω_i, also a subset of the n_{i-1} nodes that could coalesced at event i
β	the concentration parameter of Dirichlet process
G_0	the base distribution of Dirichlet process
G	a distribution over mixtures drawn from a Dirichlet process: $G \sim \text{DP}(\beta, G_0)$
u_i	the i^{th} mixture component of Dirichlet process mixture models
Λ	the covariance matrix of Brownian Diffusion
$\mathbf{I}$	the identity matrix
$\mathbb{I}$	the indicator
$\mathcal{N}$	the Gaussian distribution

2 DPMM with Gaussian Base Distribution

This section reviews how we select the restriction set Ω_i by Dirichlet process mixture models (DPMM).

Given the Brownian diffusion kernel, a natural choice for the base distribution of the DP in the DPMM is a Gaussian. We review Gibbs sampling for this model [1], which provides distributions over partitions that become the restriction set.

We initialize partitions randomly and then repeatedly resample which partition each node is in. This is possible through the exchangeablility of the Dirichlet process.

Let x_n be the current node and $\mathbf{x}_{-n}$ all other nodes, z_n the current node's cluster assignment, $\mathbf{z}_{-n}$ all other nodes' cluster assignments, n_k is the number of nodes assigned to cluster k, and N is the total number of observations. As before, β is the Dirichlet process concentration parameter. We assume that the base distribution G_0 is a Gaussian distribution with mean μ_0 and covariance Σ_0 and that each cluster has known covariance Σ_k, thus the conditional distribution is

$$p(z_n = k|\mathbf{z}_{-n}, \mathbf{x}, \mu, \beta) = \begin{cases} \frac{n_k \mathcal{N}(x_n; \hat{\mu}_k, \hat{\Sigma}_k)}{\beta+N-1} & k \text{ is old} \\ \frac{\beta \mathcal{N}(x_n; \hat{\mu}_k, \hat{\Sigma}_k)}{\beta+N-1} & k \text{ is new,} \end{cases} \tag{1}$$

where

$$\hat{\mu}_k = \frac{\mu_0 \Sigma_0^{-1} + \sum_{i \neq n} \mathbb{I}\left[z_i = k\right] x_i \cdot \Sigma_k^{-1}}{\Sigma_0^{-1} + \sum_{i \neq n} \mathbb{I}\left[z_i = k\right] \cdot \Sigma_k^{-1}}, \quad \hat{\Sigma}_k = \frac{1 + \Sigma_0^{-1}\Sigma_k + \sum_{i \neq n} \mathbb{I}\left[z_i = k\right]}{\Sigma_0^{-1} + \sum_{i \neq n} \mathbb{I}\left[z_i = k\right] \cdot \Sigma_k^{-1}}$$

This is also called the infinite Gaussian mixture model (IGMM) [2], which clusters nodes with similar feature values, providing useful candidates for the coalescent to merge.

3 Generating Synthetic Data

To test how well the different methods capture hierarchical data, we generate synthetic hierarchical data with a known structure and test whether our model can recover the hierarchy. According to Berestycki [3], given n_{i-1} nodes at time t_{i-1} and $t_i = t_{i-1} - \delta_i$, the expected number of nodes that merge at time t_i is

$$1 + \delta_i \left(\sum_{k_i=2}^{n_{i-1}} (k_i - 1) \binom{n_{i-1}}{k_i} \lambda_{n_{i-1}}^{k_i} \right). \tag{2}$$

Therefore we start with n_0 nodes, sample a duration time δ_i, and compute the expected number of nodes to be merged at time t_i; we then merge that number of nodes and repeat until there is only one node.

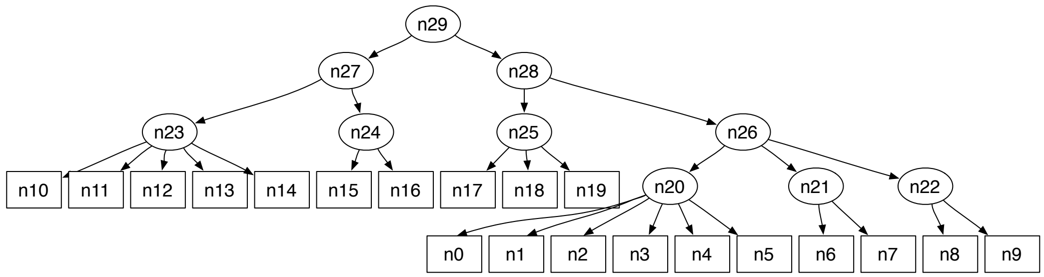
Next we generate the features for nodes from a Gaussian kernel. We start with the root node as a multivariate Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$, where the mean $\mu_0 = (0, \cdots, 0)$ and $\Sigma_0 = \rho_0 \mathbf{I}$ ($\mathbf{I}$ is the identity matrix). For each child, we sample the feature vector y_c from the parent Gaussian $\mathcal{N}(y_p, \Sigma_p)$, and set $\Sigma_c = \frac{1}{n}\rho_p \mathbf{I}$. In this experiment, we generate the data with parameter $\rho_0 = 10$. Labels are assigned based on the root's children; each subtree rooted at a child of the root receives the same label. This class label is used to calculate the metrics defined above.

4 Synthetic Trees

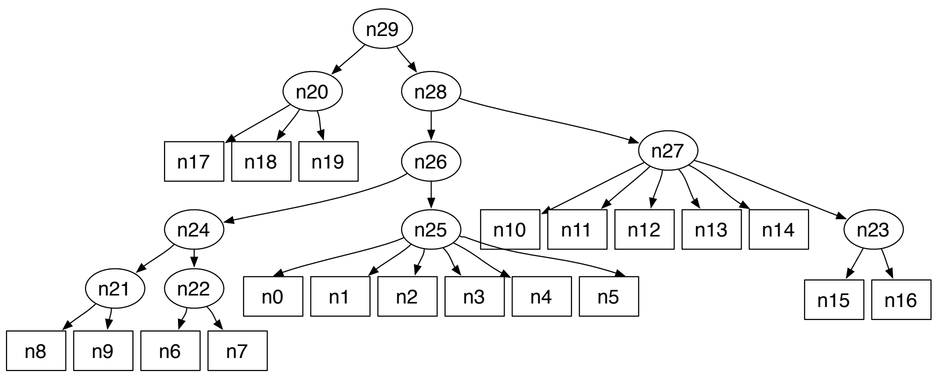
This section compares the constructed synthetic trees of Beta coalescent and Kingman's coalescent with the true synthetic trees. For all the following trees, the square nodes are the observed leaf nodes, and the circle nodes are the detected hidden internal nodes.

4.1 Tree1: n = 20

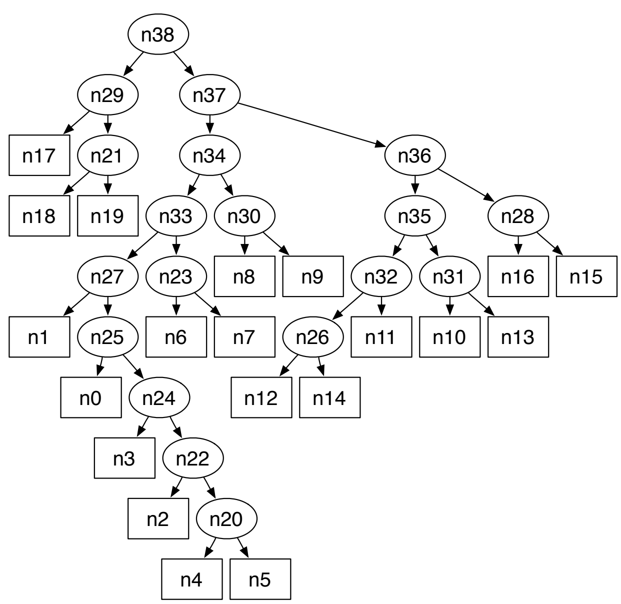
- True synthetic tree



- Constructed tree from Beta coalescent

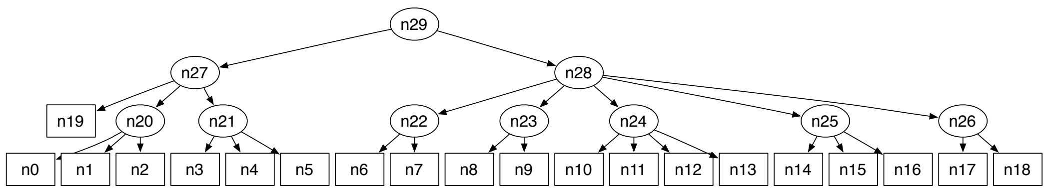


- Constructed tree from Kingman's coalescent

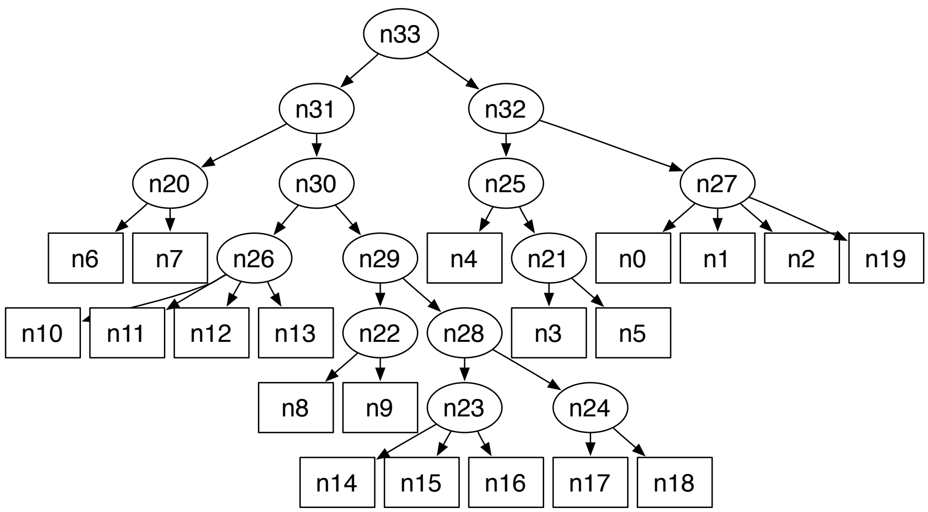


4.2 Tree2: n = 20

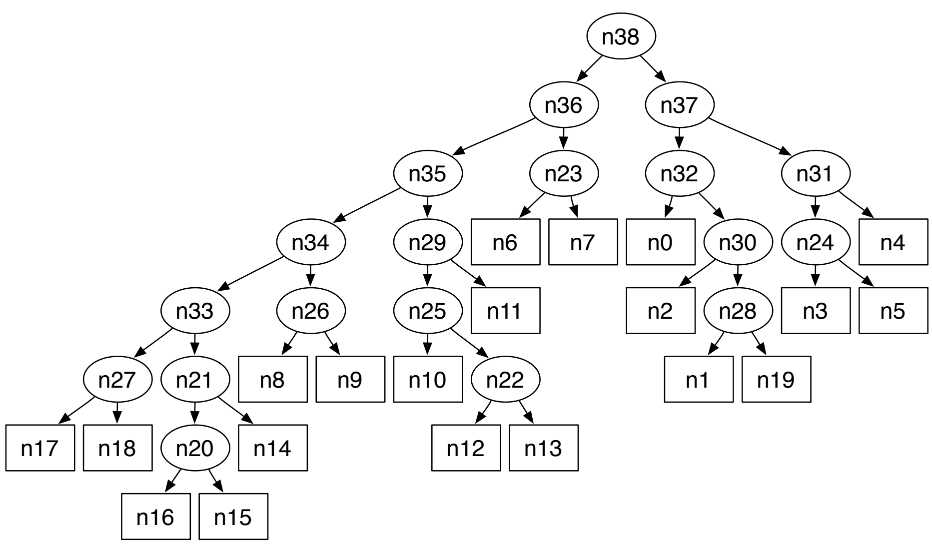
- True synthetic tree



- Constructed tree from Beta coalescent

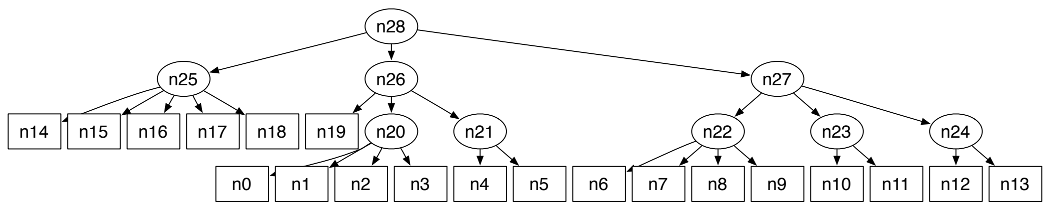


- Constructed tree from Kingman's coalescent

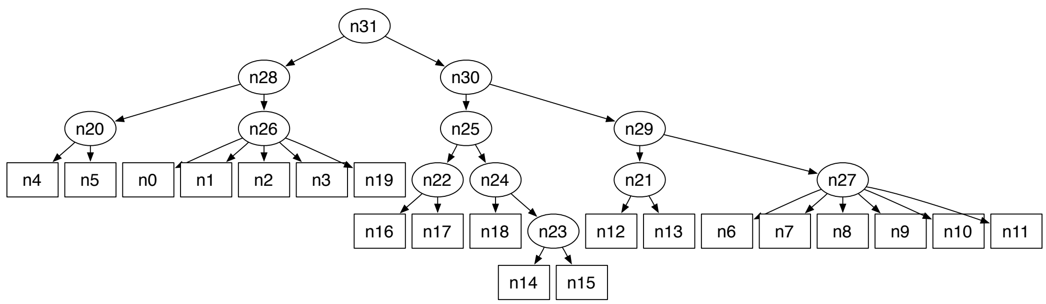


4.3 Tree3: n = 20

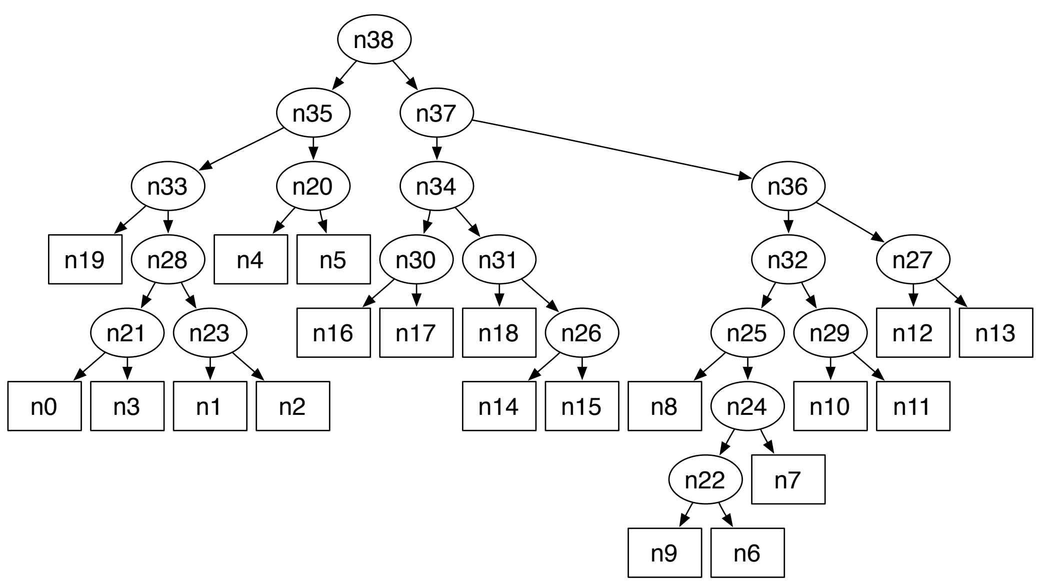
- True synthetic tree



- Constructed tree from Beta coalescent

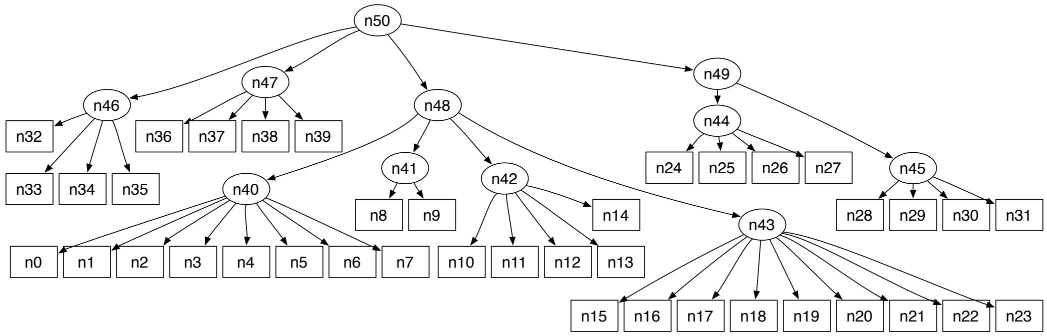


- Constructed tree from Kingman's coalescent

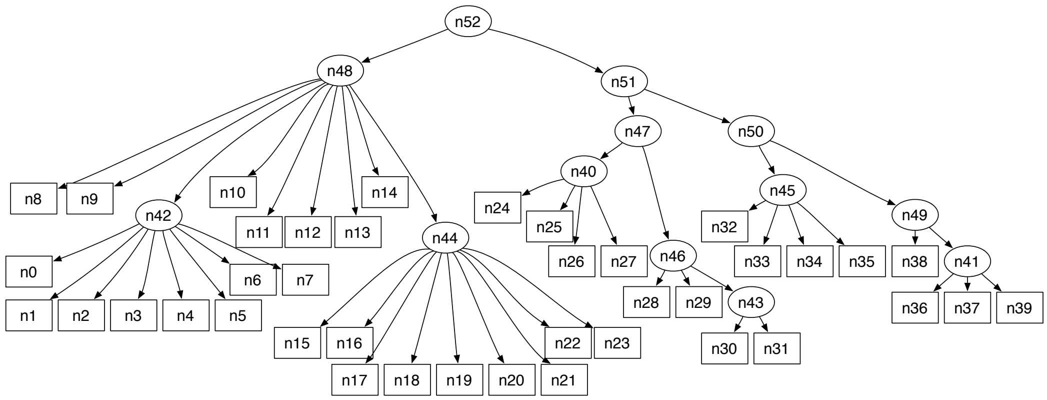


4.4 Tree4: n = 40

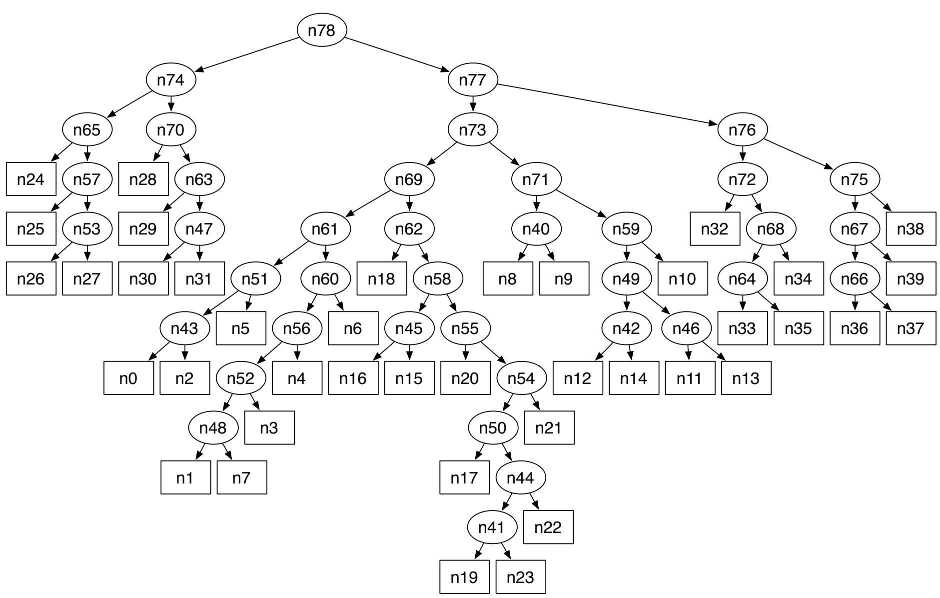
- True synthetic tree



- Constructed tree from Beta coalescent

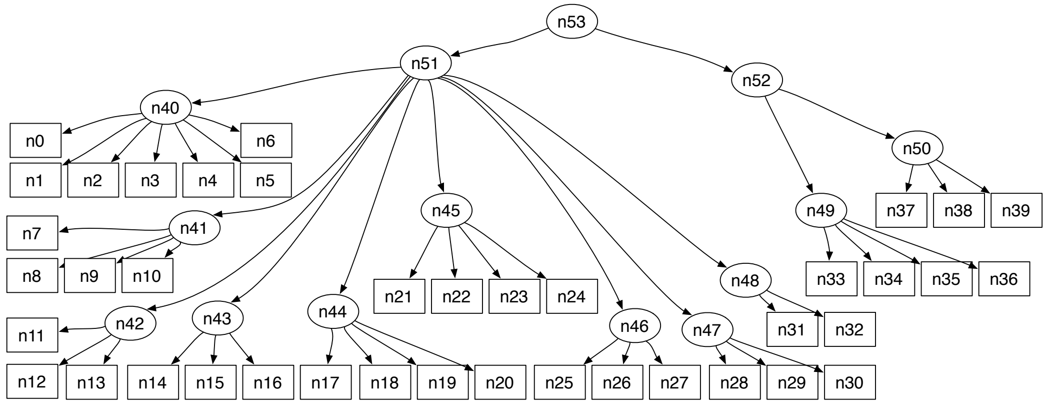


- Constructed tree from Kingman's coalescent

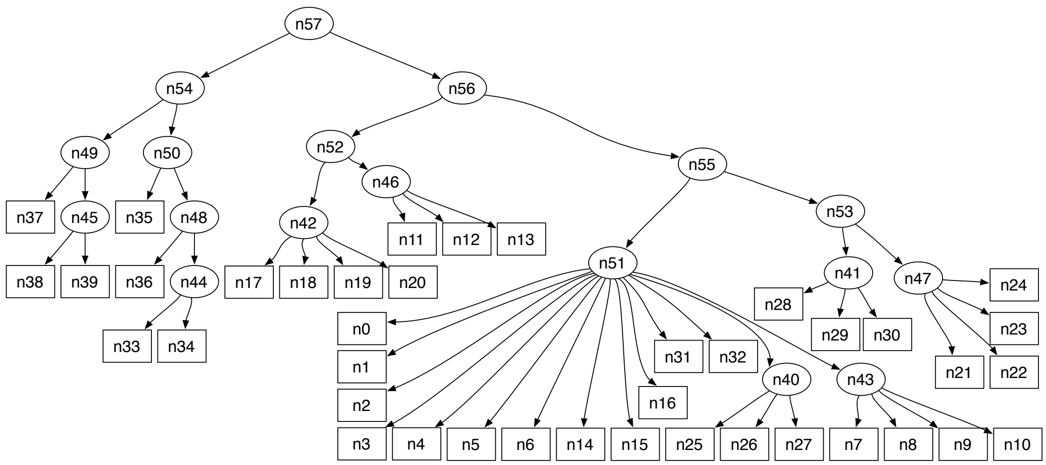


4.5 Tree5: n = 40

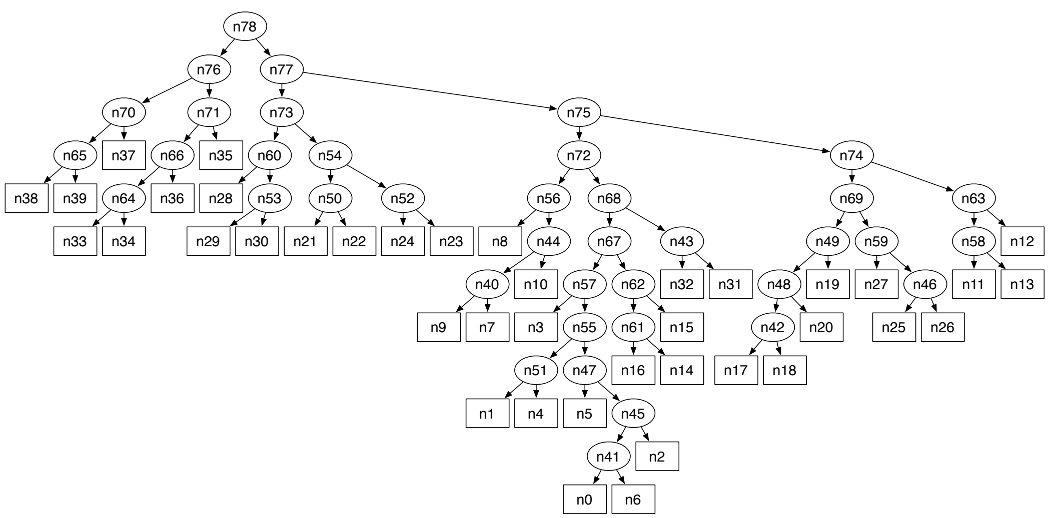
- True synthetic tree



- Constructed tree from Beta coalescent



- Constructed tree from Kingman's coalescent



References

[1] Neal, R. M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[2] Rasmussen, C. E. The infinite Gaussian mixture model. In *NIPS*. 2000.

[3] Berestycki, N. Recent progress in coalescent theory. In *Ensaios Matematicos*, vol. 16. 2009.