MCMC based machine learning

(Bayesian Model Averaging) [a].

Nicos Angelopoulos

n.angelopoulos@ed.ac.uk

School of Biological Sciences

Biochemistry Group

University of Edinburgh, Scotland, UK.

[a] Collaborative work with James Cussens, York University, jc@cs.york.ac.uk

MCMC Overview

Class of sampling algorithms that estimate a posterior distribution.

Markov chain
construct a chain of visited values, $M_1, M_2, \ldots, M_n$, by proposing M_* from M_i , with probability $q(M_*, M_i)$. Use prior knowledge, $p(M_*)$ and relative likelihood of the two values, $p(D|M_*)/p(D|M_i)$ to decide chain construction.

Monte Carlo
Use the chain to approximate the posterior $p(M|D)$.

Bayesian learning with MCMC

Given some data D and a class of statistical models $\mathcal{M}$ ($M \in \mathcal{M}$) that can express relations in the data, use MCMC to approximate normalisation factor in Bayes' theorem

$$p(M|D) = \frac{p(D|M)p(M)}{\sum_M p(D|M)p(M)}$$

$p(M)$ is the prior probability of each model

$p(D|M)$ the likelihood (how well the model fits the data)

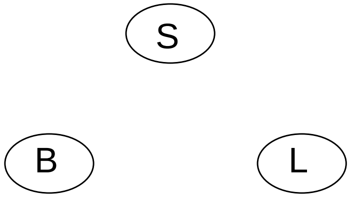
$p(M|D)$ the posterior

Example: Data

	smoker	bronchitis	l_cancer
person 1	y	y	n
person 2	y	n	n
person 3	y	y	y
person 4	n	y	n
person 5	n	n	n

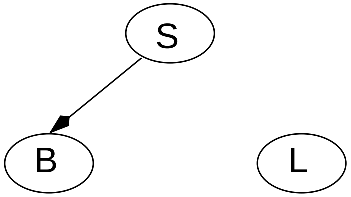
Example: Models

B_1



[b-[],l-[],s-[]]

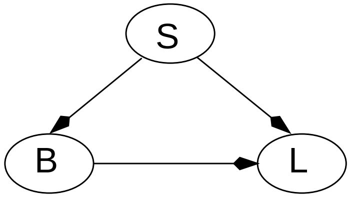
B_2



[b-[s],l-[],s-[]]

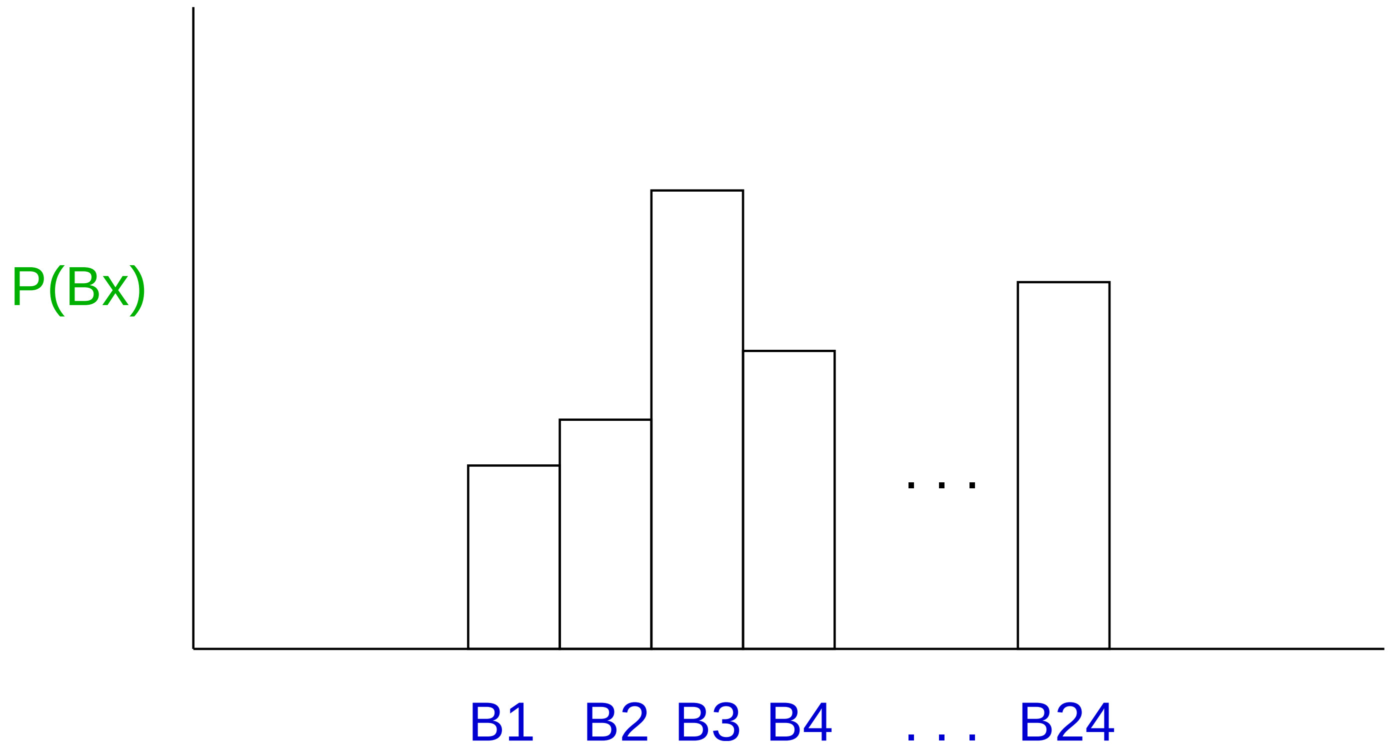
$\ldots$

B_{24}



[b-[s],l-[b,s],s-[]]

Example: Objective



$$\sum_{B_x} p(B_x) = 1$$

Metropolis-Hastings (M-H) MCMC

0. Set $i = 0$ and find M_0 using the prior.
1. From M_i produce a candidate model M_*. Let the probability of reaching M_* be $q(M_*, M_i)$.
2. Let

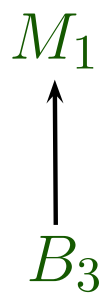
$$\alpha(M_i, M_*) = min\left\{\frac{q(M_*, M_i)P(D|M_*)P(M_*)}{q(M_i, M_*)P(D|M_i)P(M_i)}, 1\right\}$$

$$M_{i+1} = \begin{cases} M_* & \text{with probability } \alpha(M_i, M_*) \\ M_i & \text{with probability } 1 - \alpha(M_i, M_*) \end{cases}$$

3. If i reached limit then terminate, else set $i = i + 1$ and repeat from 1.

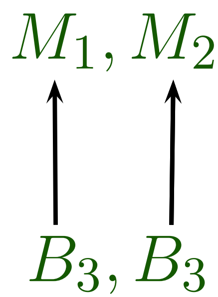
Example: MCMC

Markov Chain:



Example: MCMC

Markov Chain:



Example: MCMC

Markov Chain:

$$\begin{array}{cccccc} M_1, & M_2, & M_3, & M_4, & M_5, & \ldots \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \\ B_3, & B_3, & B_{10}, & B_3, & B_{24}, & \ldots \end{array}$$

Example: MCMC

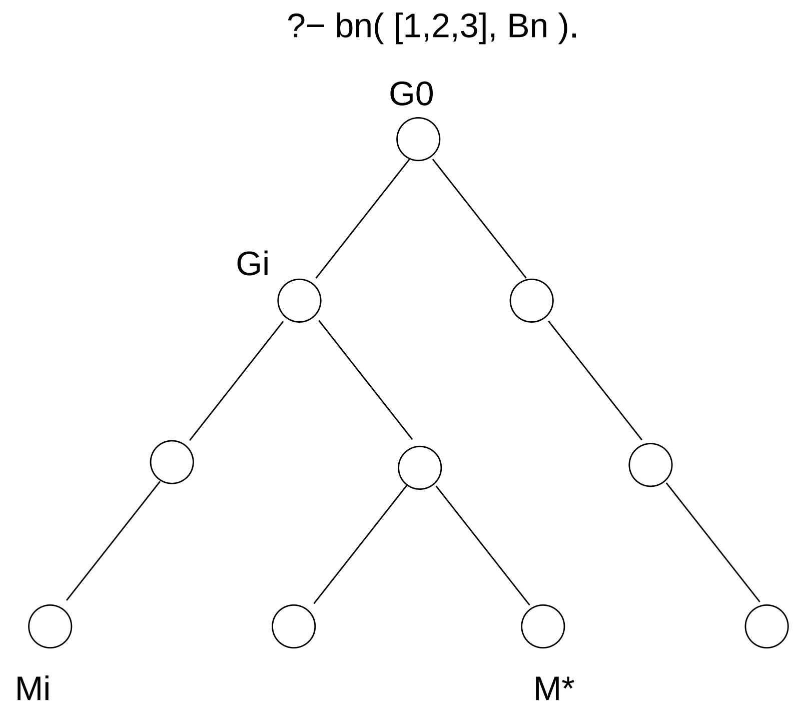
Markov Chain:

$$\begin{array}{cccccc} M_1, & M_2, & M_3, & M_4, & M_5, & \ldots \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \\ B_3, & B_3, & B_{10}, & B_3, & B_{24}, & \ldots \end{array}$$

Monte Carlo:

$$p(B_k) = \frac{\#(B_k)}{\sum_{B_x} \#(B_x)}$$

SLP defined model space



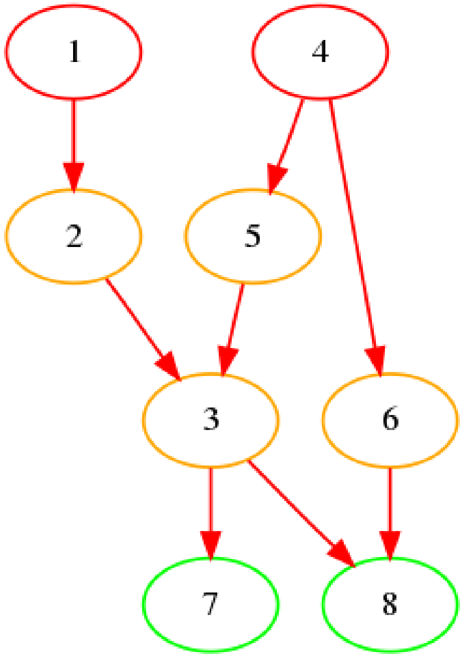
From M_i identify G_i then sample forward to $M_\star$.
$q(M_i, M_\star)$ is the probability of proposing $M_\star$ when M_i is the current model.

BN Prior

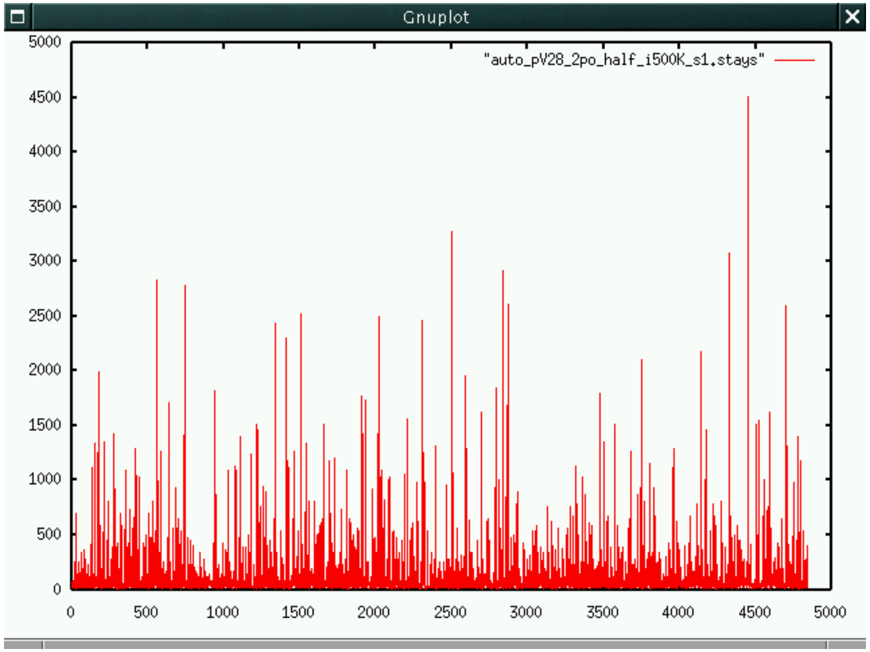
```
bn( OrdNodes, Bn ) :-
    bn( Nodes, [], Bn ).

bn( [], _PotPar, [] ).
bn( [H|T], PotPar, [H-SelParOfH|RemBn] ) :-
    select_parents( PotPar, H, SelParOfH ),
    bn( T, [H|PotPar], RemBn ).

select_parents( [], [] ).
select_parents( [H|T], Pa ) :-
    include_element( H, Pa, RemPa ),
    select_parents( T, TPa ).

1/2 : include_element(  H, [H|TPa], TPa ).
1/2 : include_element( _H,    TPa,   TPa ).
```

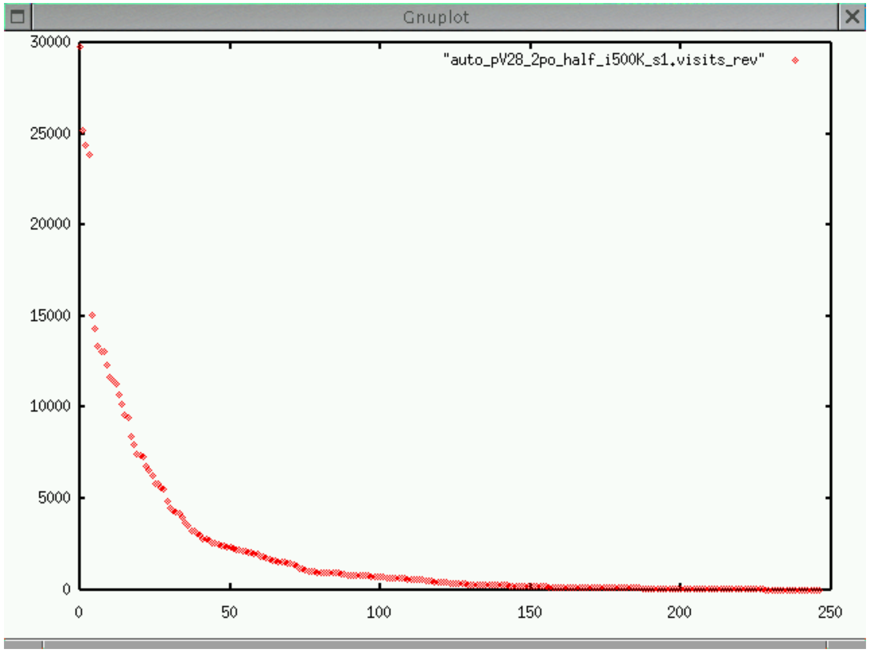
example BN (Asia)

For example ?- bn([1,2,3,4,5,6,7,8], M).
M = [1-[],2-[1],3-[2,5],4-[],5-[4],6-[4],7-[3],8-[3,6]].



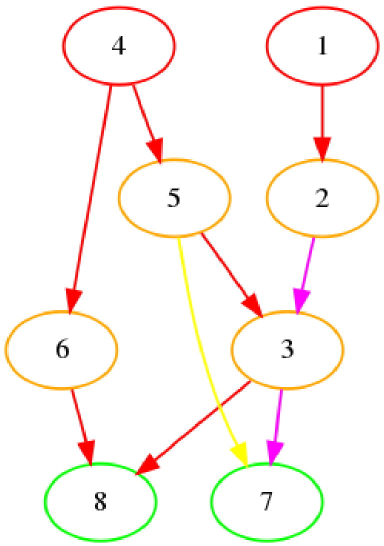
visits and stays





Edges recovery

With topological ordering constraint and a maximum of 2 parents per node, the algorithm recovers most of the BN arcs in 0.5 M iterations. For example for a .99 cut-off we have :



Missing :

- $2 \longrightarrow 3$ (.84)
- $3 \longrightarrow 7$ (.47)

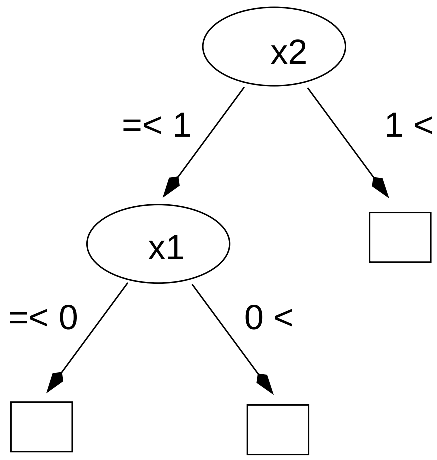
Superfluous :

- $5 \longrightarrow 7$

CART priors

$$P_{\mathrm{split}}(\eta) = \alpha(1 + d_\eta)^{-\beta}$$

?- cart(M).



M = node(b, 1, node(a,0,leaf,leaf), leaf)

```
1 - Sp: [Sp]: cart( Data, D, A/B, leaf(Data) ).
    Sp: [Sp]: cart( Data, D, A/B, node(F,V,L,R) ) :-
         branch( Data, F, V, LData, RData ),
         D1 is D + 1,
         NxtSp is A * ((1 + D1) ^ -B),
         [NxtSp] : cart( LData, D1, A/B, L ),
         [NxtSp] : cart( RData, D1, A/B, R ).
```

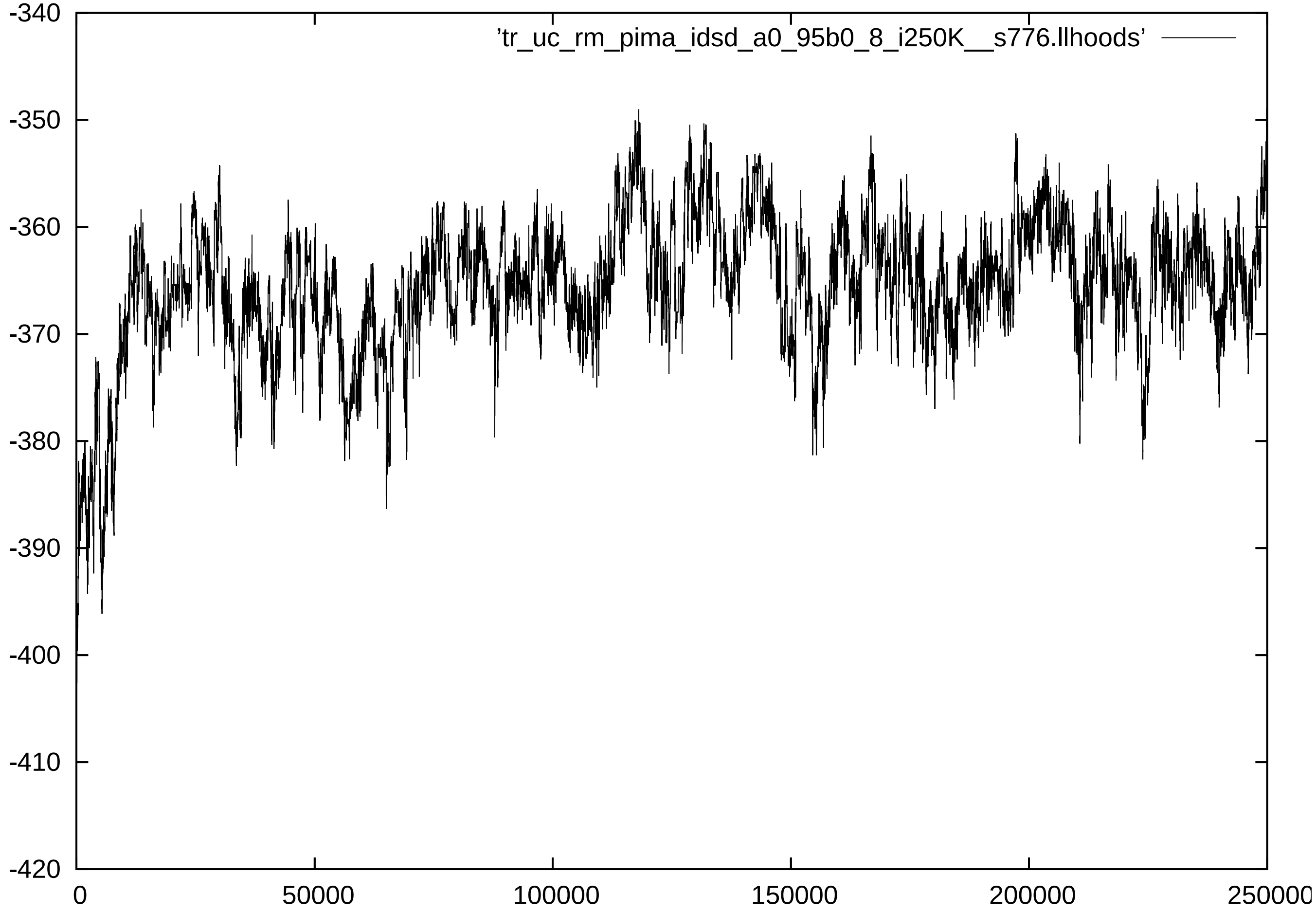
Experiment

Pima Indians Diabetes Database
768 complete entries of 8 variables.

Denison et.al. run 250,000 iterations of local perturbations.
Their best likelihood model: -343.056

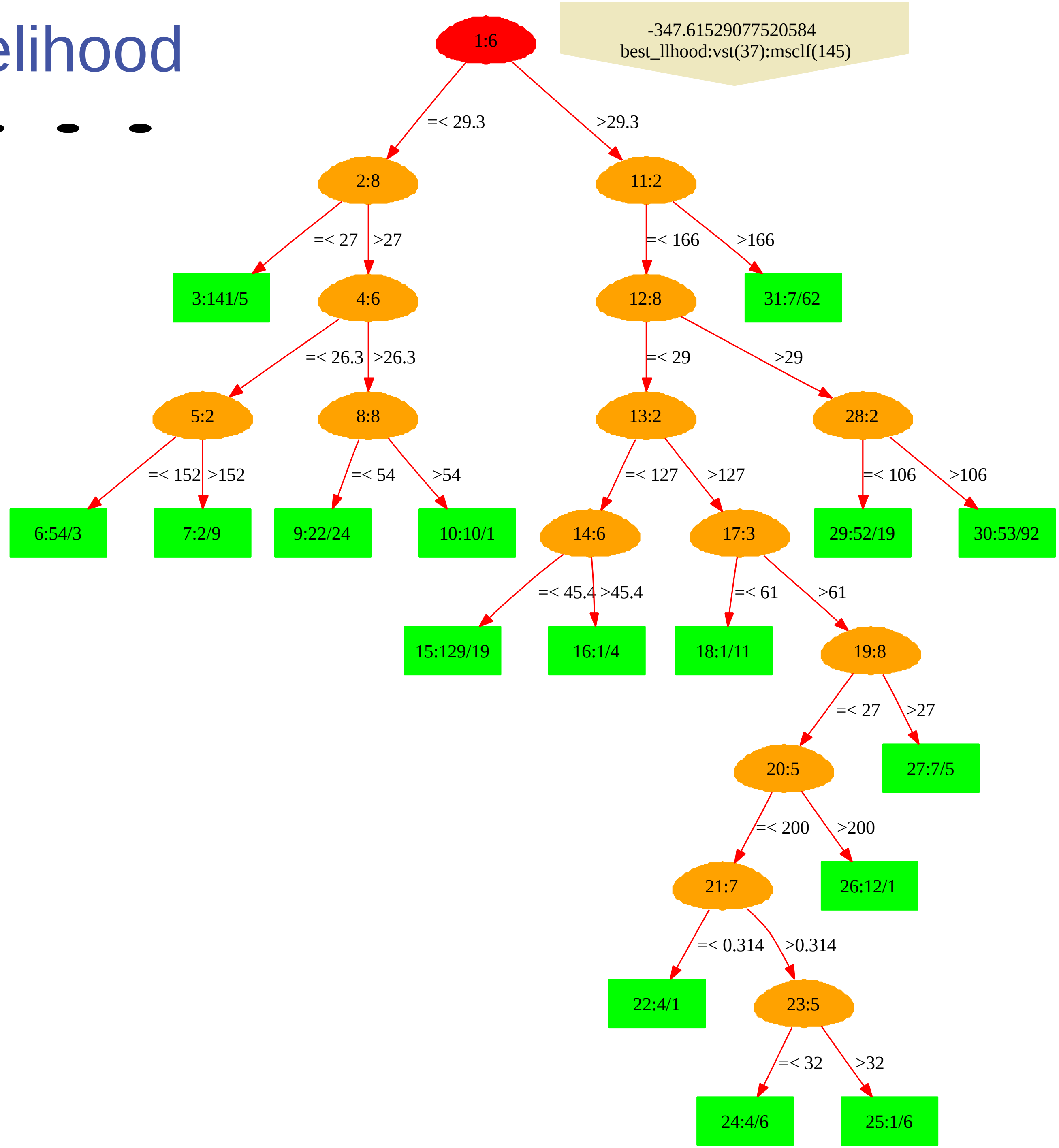
Our experiment run for 250,000 iterations with branch replacing.
Parameters: uniform choice proposal, $\alpha = .95$ $\beta = .8$
Our best likelihood model: -347.651

Likelihoods trace



$$\beta = .8, \alpha = .95, proposal = uniform\ choice$$

Best likelihood



in Kyoto

—— . . .

Models: HMRFs for clustering.

Likelihood: design and implement a likelihood-ratio function for HMRFs.

Proposal: implement function(s) for reaching proposal model.

Application: to real data.

SLPs: for more complex priors.