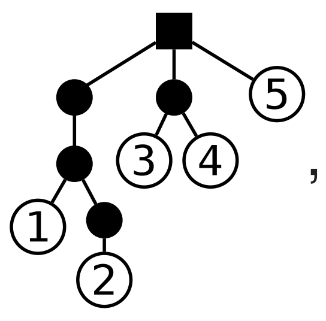
Improved importance sampling of phylogenies

Mathias C. Cronjäger (cronjager@stats.ox.ac.uk),
Paul Jenkins (Warwick), Jotun Hein (Oxford)

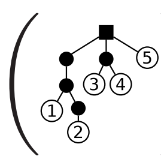
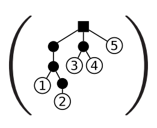
Motivation: likelihoods with an unknown ancestral tree

Given a set of aligned sequences, e.g.

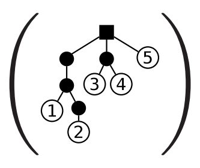
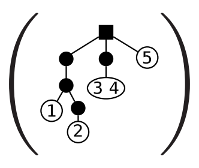
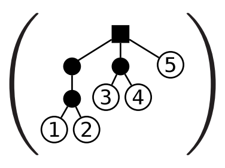
Sequence 1: … T … G … A … A …
Sequence 2: … T … G … A … G …
Sequence 3: … A … A … T … A …
Sequence 4: … A … A … T … A …
Sequence 5: … A … A … A … A …

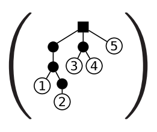
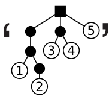
represented as ,

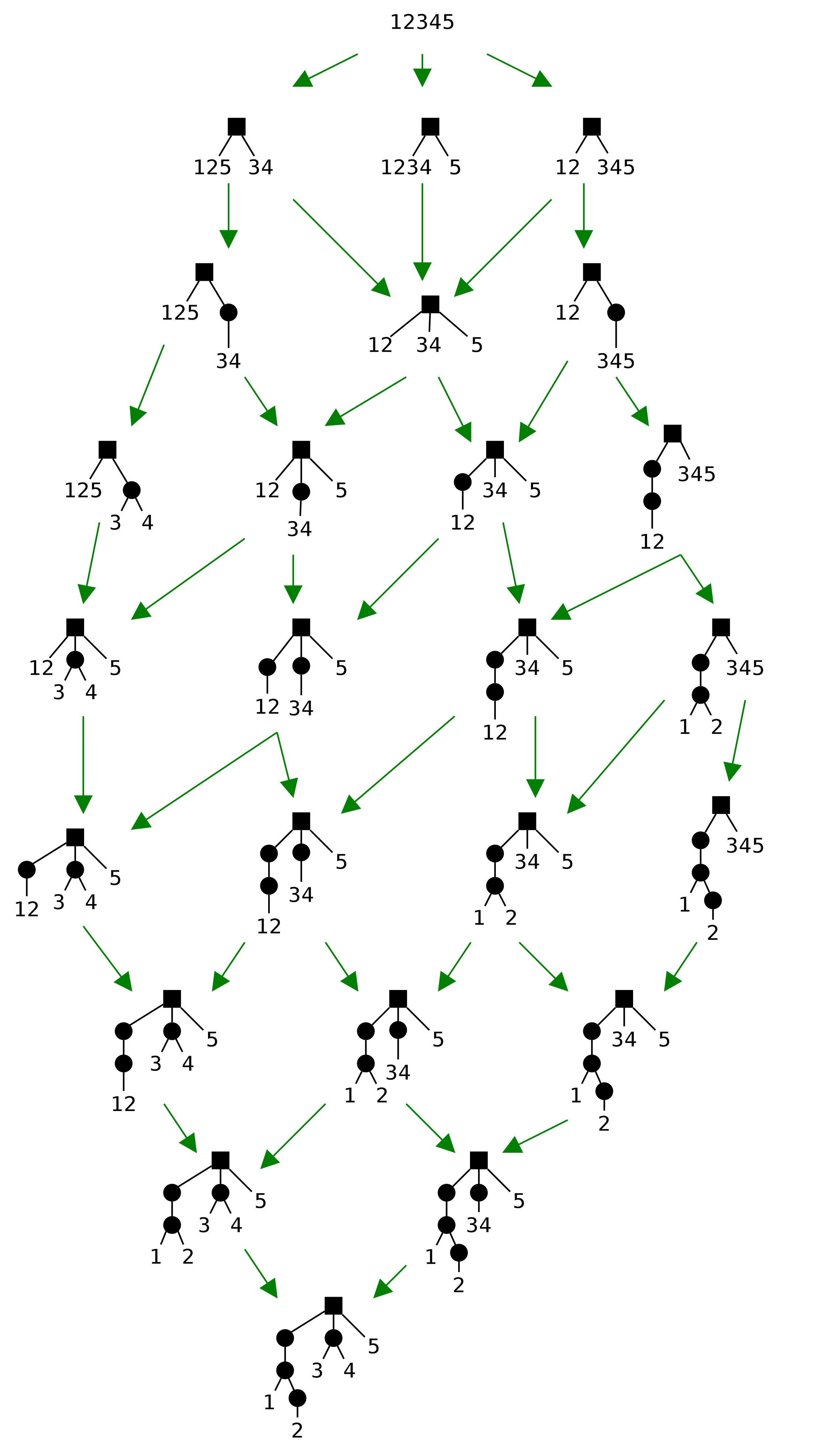
the probability of having evolved from some initial sequence (here $\ldots A \ldots A \ldots A \ldots A \ldots$) may be expressed:

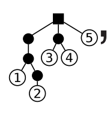
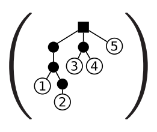
$\mathbb{P}\Big($  $\Big) = \sum_{x \in \text{Histories}(\ldots)}$  $\mathbb{P}(x)$

which in turn may be expressed by conditioning on the most recent event:

$\mathbb{P}\Big($  $\Big) = \frac{1}{\theta+4}\mathbb{P}\Big($  $\Big) + \frac{\theta}{\theta+4}\frac{1}{5}\mathbb{P}\Big($  $\Big).$

If we apply this conditioning-trick recursively, computing $\mathbb{P}\Big($  $\Big)$ reduces to computing a weighted sum over all paths from "12345" to "" in the *ancestral graph* below:

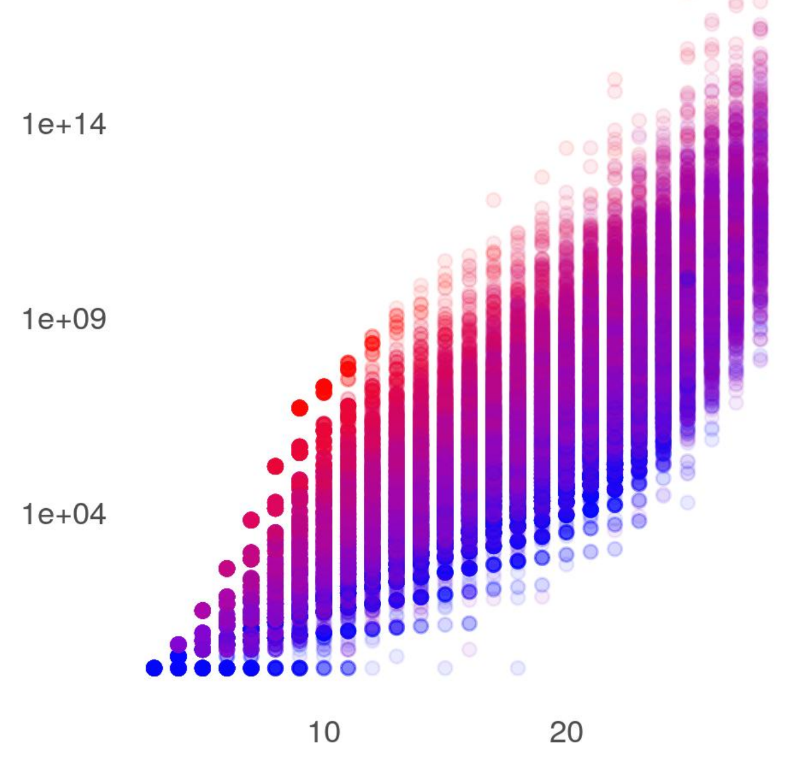


- number of nodes = number of distinct terms in $\mathbb{P}$-recursion,
- number of paths "$12345 \to \ldots \to$ " = number of execution paths when evaluating $\mathbb{P}\Big($  $\Big)$ via tail-recursion (without memoization/tabling).

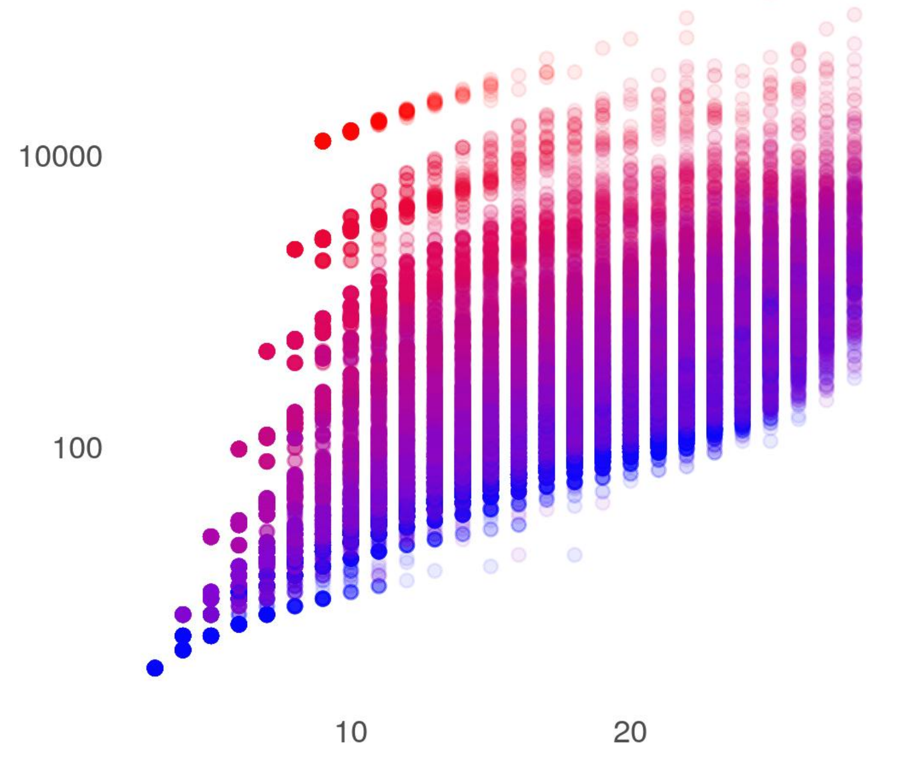
Challenge: A growing graph of ancestral states

As the number of sequences n and segregating sites s increases, it quickly becomes computationally intractable to recursively compute exact likelihoods.

Rank (n + s) vs. #paths
colour = max degree (blue=low, red=high)

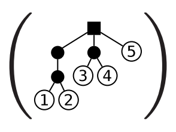


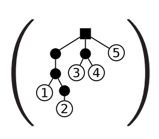
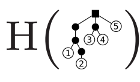
Rank (n+s) vs. #past states
colour = max degree (blue=low, red=high)

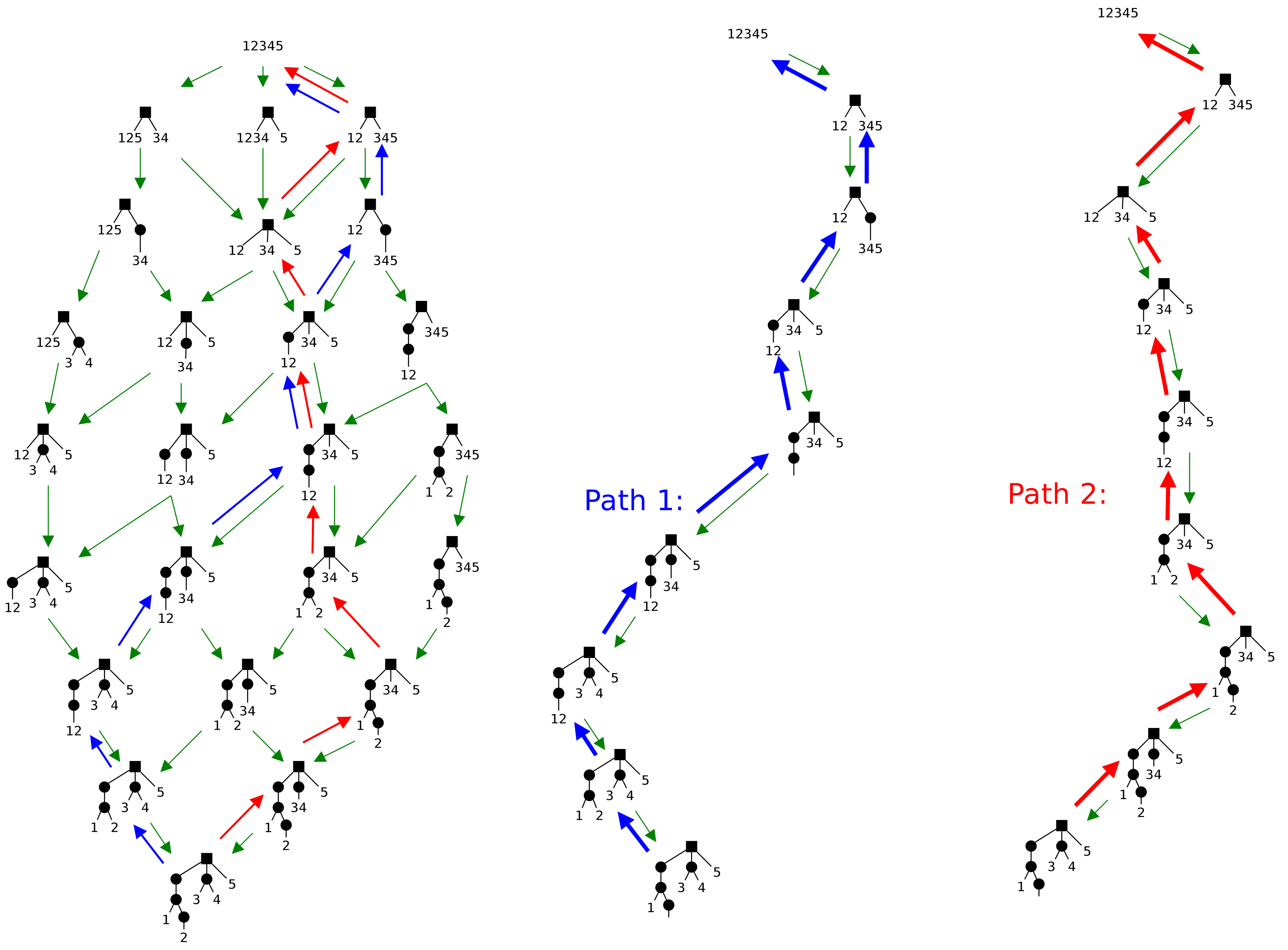


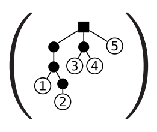
We need methods which do not pre-suppose the ancestral graph, since it is a priori unknown and generating it is as hard as computing likelihoods.

Importance Sampling of ancestral paths

We can approximate probabilities of aligned sequences–e.g. $\mathbb{P}\Big($  $\Big)$–by sampling ancestral histories $X_1, \ldots, X_N \overset{iid}{\sim} \mathbb{Q} \ll \mathbb{P}$ and relying on the following approximation:

$\mathbb{P}\Big($  $\Big) = \sum_{x \in \mathrm{H}(\ldots)}$  $\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}\mathbb{Q}(x) = \mathbb{E}_{X\sim\mathbb{Q}}\left[\frac{\mathbb{P}(X)}{\mathbb{Q}(X)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{P}(X_i)}{\mathbb{Q}(X_i)}$



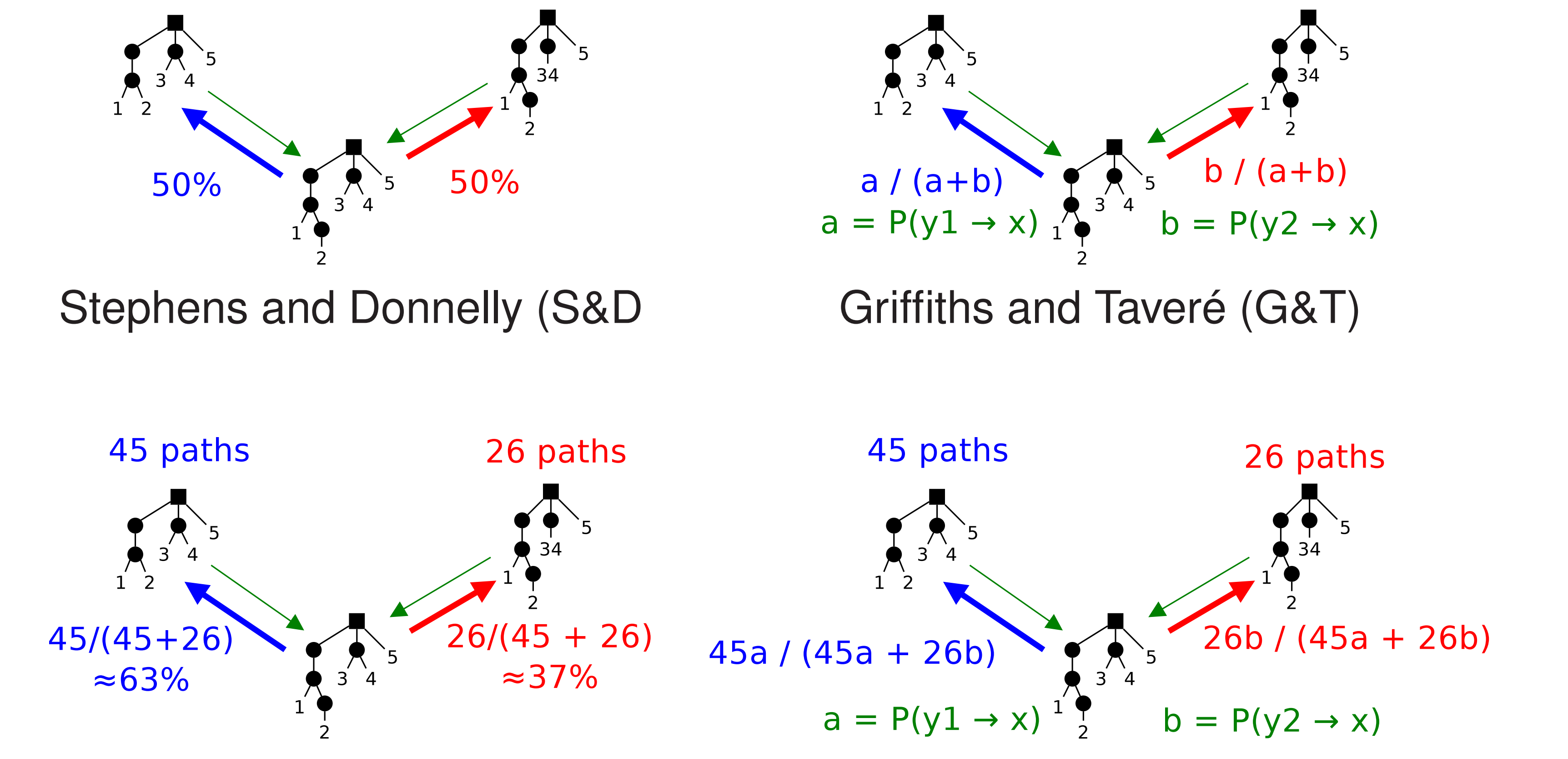
e.g. $\mathbb{P}\Big($  $\Big) \approx \frac{1}{2}\left(\frac{\mathbb{P}(\text{Path 1})}{\mathbb{Q}(\text{Path 1})} + \frac{\mathbb{P}(\text{Path 2})}{\mathbb{Q}(\text{Path 2})}\right)$

For this approach to work effectively, $\mathbb{Q}$ should satisfy:

1. $\mathbb{Q}$ must approximate $\mathbb{P}$ well on the space of histories;
2. sampling $X_i \sim \mathbb{Q}$ should be fast;
3. computing $\mathbb{Q}(X_i)$ should be fast.

Sequential sampling schemes

Existing proposal distributions are all sequential: they construct paths step-by step from the bottom up. They differ by how the next step in a path is sampled.



Stephens and Donnelly (S&D

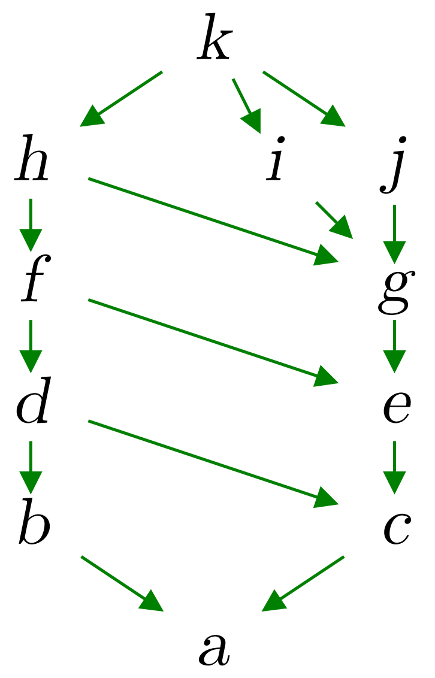
Griffiths and Taveré (G&T)

simple combinatorial sampling

G&T + combinatorial correction

Path density bias and path counting

Any step-by-step scheme which does not penalize choices which "lead to fewer choices down the line", will be biassed in favour of low-density regions of path-space, e.g.

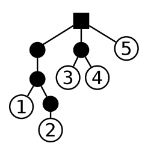


Path	Weight (S&D)
khfdba	*1/2*
khfdca	*1/4*
khfeca	*1/8*
khgeca	*1/24*
kigeca	*1/24*
kjgeca	*1/24*

To correct for path density bias, we must be able to count ancestral histories effectively (i.e. without generating the ancestral graph), which we do as follows:

$$h(T) = \sum\nolimits_{S \subsetneq [r], 1 \in S} h(\{T_i \mid i \in S\})h(\{T_i \mid i \notin S\})\binom{(\sum_{i=1}^{r} k_i) - 2}{(\sum_{i \in S} k_i) - 1}$$

whereby we here encode rooted unordered trees as nested systems of sets, e.g.

 $= \{\{\{1, \{2\}\}\}, \{3, 4\}, 5\}.$