Mixing Times for the Branch-Rotation Chain on Cladograms (or the Triangulation Walk)

This concerns two problems which are similar (intuitively, at least). Consider the regular n-gon. There are a finite number $C_n = \frac{2n-4)!}{(n-1)!(n-2)!}$ of *triangulations*, that is ways to draw diagonals which partition the n-gon into triangular regions (C_n is a Catalan number: see [5] pages 219–229). One can define a discrete time Markov chain on the space of triangulations of the n-gon as follows. In each step

> pick uniformly at random a diagonal line; delete it, to leave a quadrilateral; then insert the opposite diagonal of that quadrilateral to get a new triangulation.

A different combinatorial set is the set of *n-cladograms*. Such a cladogram, illustrated in figure 1, has leaves labeled $1, 2, \ldots, n$, an unlabeled root (at the top) and binary splits, where we do not distinguish left and right subtrees. (Cladograms are one formalization of *phylogenetic trees* from biological systematics, indicating evolutionary relationships between species. The number of n-cladograms equals $\frac{(2n-2)!}{2^{n-1}(n-1)!}$.) One can define several Markov chains on the set of n-cladograms (see [1, 4] for a more easily-analyzed chain), but the following type of chain seems most interesting.

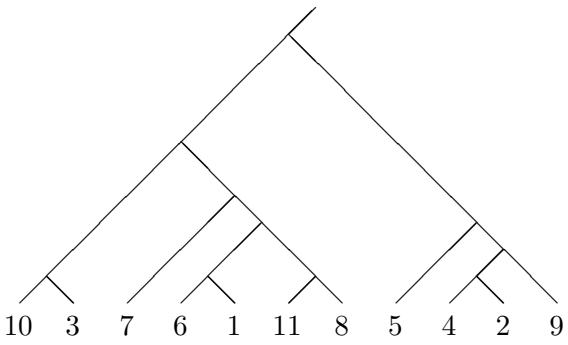

Figure 1. A cladogram on 11 species.

A n-cladogram has $2n-1$ edges. Pick one edge (not the edge at the root) uniformly at random; in figure 1, say we pick the edge upwards from the common ancestor of $\{11, 8\}$. Cut this edge at its top, thus separating the 3-edge subcladogram on $\{11, 8\}$ and making the two other edges at the cut-

point merge into a single edge e from the common ancestor of $\{6, 1\}$ to the common ancestor of $\{7, 6, 1\}$. Now there are exactly 4 edges adjacent to e, viz the edges leading upwards from

$$6, \quad 1, \quad 7, \quad \text{the common ancestor of } \{7, 6, 1\}.$$

Pick each of these 4 edges with chance 1/4, and reattach the subcladogram to the middle of that edge. If we picked the edge upwards from the common ancestor of $\{7, 6, 1\}$, then we would obtain the cladogram in figure 2. (In general e might have less than 4 adjacent edges, in which case with the remaining probability we make no change).

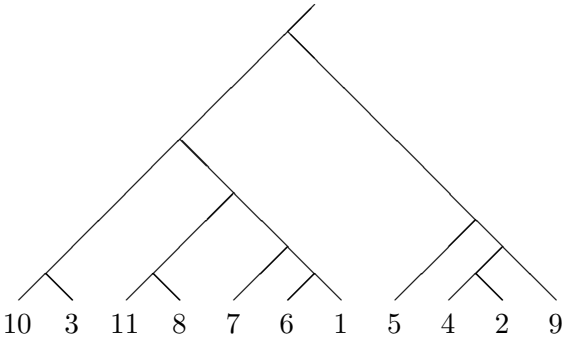

Figure 2. A step of the chain from figure 1.

Being reversible, each chain has largest eigenvalues $1 = \lambda_1 > \lambda_2$, and the *relaxation time* defined as $1/(1 - \lambda_2)$ has an interpretation as a mixing time parameter. In each chain it is easy to show (by the usual technique of applying the variational characterization of λ_2 to a suitable test function) that the relaxation is at least order $n^{3/2}$ as $n \to \infty$.

PROBLEM. For each chain, show that the relaxation time is at most order $n^{3/2}$.

Discussion

There are good heuristic reasons (too lengthy to explain here) for expecting these two chains to have similar behavior. The chain on triangulations is discussed in [2, 3], who obtain an $O(n^4)$ upper bound. Over the last 20 years, techniques has been developed which enable one to find the correct order of magnitude of the mixing times for natural random walks on familiar combinatorial structures; this "triangulation walk" example is perhaps the simplest structure for which correct order is unproved. It seems naturalk to try the "distinguished paths" method, but it seems surprisingly hard to take

the first step of defining a "natural" path in tree-space between an arbitrary pair of cladograms.

The "cladograms" chain has a semi-applied story. Reconstructing phylogenetic trees from actual biological data is a large-scale academic activity; it involves algorithmically hard optimization problems which in practice are attacked via heuristic "local search" methods, exploring the space of cladograms to find a "best fit" to data. One class of algorithms uses MCMC (Markov chain Monte Carlo), built over a "base chain" like ours. The data-dependent chains which arise in practice are so complicated that rigorous theoretical analysis seems hopeless, but understanding the base "no data" chain is a natural first step.

History. I have posed this in talks and conversation since around 2000, originating from [1].

References

[1] D.J. Aldous. Mixing time for a Markov chain on cladograms. *Combin. Probab. Comput.*, 9:191–204, 2000.

[2] L. McShine and P. Tetali. On the mixing time of the triangulation walk and other Catalan structures. In *Randomization Methods in Algorithm Design (Princeton NJ 1997)*, number 43 in DIMACS Ser. Discrete Math. Theoret. Comput. Sci., pages 147–160. Amer. Math. Soc., 1999.

[3] M. Molloy, B. Reed, and W. Steiger. On the mixing time of the triangulation walk. In *Randomization Methods in Algorithm Design (Princeton NJ 1997)*, number 43 in DIMACS Ser. Discrete Math. Theoret. Comput. Sci., pages 179–190. Amer. Math. Soc., 1999.

[4] J. Schweinsberg. An $O(n^2)$ bound for the relaxation time of a Markov chain on cladograms. Technical Report 572, Statistics Dept., U.C. Berkeley, 2000.

[5] R.P. Stanley. *Enumerative Combinatorics*, volume 2. Cambrdige University Press, 1999.