Maximal clades in random binary search trees

Svante Janson

A phylogenetic tree, or a full binary tree is a tree where every node has outdegree 0 or 2; nodes with outdegree 0 are called external and nodes with outdegree 2 internal. By eliminating all external nodes, we get a binary tree, and this yields a bijection between phylogenetic trees with \( n + 1 \) external nodes and binary trees with \( n \) nodes.

The clade of an external node \( v \) in a phylogenetic tree is the set of external nodes that are descendants of the parent of \( v \). (This is called a minimal clade by [1] and [2].) Note that two clades are either nested or disjoint, and that the set of maximal clades forms a partition of the set of external nodes. We let \( F(T) \) denote the number of maximal clades of a phylogenetic tree \( T \). The maximal clades, and the number of them, were introduced by [4], together with a biological motivation, and further studied by [3].

Translated to the corresponding binary tree (i.e., the internal nodes), a clade is thus a node having outdegree at most 1, and a maximal clade is a clade such that all ancestors have outdegree 2.

We consider a random binary search tree \( T_n \) (which corresponds to the Yule–Harding model of a random phylogenetic tree) and the number of maximal clades \( X_n := F(T_n) \) in it. We consider asymptotics as \( n \to \infty \).

It was proved by [5] and [3] that

\[
E X_n = E F(T_n) = \alpha n + O(1),
\]

where that the mean number of maximal clades \( E X_n \sim \alpha n \), where

\[
\alpha = \frac{1 - e^{-2}}{4}.
\]

Moreover, [3] found also corresponding results for the variance and higher central moments:

\[
E(X_n - E X_n)^2 \sim 4\alpha^2 n \log n,
\]

and for any fixed integer \( k \geq 3 \),

\[
E(X_n - E X_n)^k \sim (-1)^k \frac{2k}{k - 2} \alpha^k n^{k-1}.
\]

As a consequence of (3)–(4), the limit distribution of \( F(T_n) \) (after centering and normalization) cannot be found by the method of moments. Nevertheless, [3] further proved asymptotic normality, where, unusually, the normalizing uses (the square root of) half the variance:

\[
X_n - E X_n \quad \frac{d}{\sqrt{2\alpha^2 n \log n}} \to N(0,1).
\]

We use probabilistic methods to reprove these theorems, together with some further results. In particular, we can explain the appearance of half the variance in (5) as follows:
Fix a sequence of numbers \( N = N(n) \), and say that a clade is small if it has at most \( N + 1 \) elements, and large otherwise. Let \( X_n^N \) be the number of maximal small clades, i.e., the small clades that are not contained in any other small clade. It turns out that a suitable choice of \( N \) is about \( \sqrt{n} \); we have for example the following.

**Theorem 1.** Let \( N := \sqrt{n} \). Then \( \text{Var}(X_n^N) \approx 2\alpha^2 n \log n \) and

\[
\frac{X_n^N - \mathbb{E}X_n^N}{\sqrt{\text{Var}X_n^N}} \xrightarrow{d} N(0,1).
\]

Furthermore, \( X_n - X_n^N = o_p\left(\sqrt{\text{Var}X_n^N}\right) \) and \( \mathbb{E}X_n - \mathbb{E}X_n^N = o\left(\sqrt{\text{Var}X_n^N}\right) \), so we may replace \( X_n^N \) by \( X_n \) in the numerator of (6). However,

\[
\text{Var}(X_n - X_n^N) \approx \text{Var}(X_n^N) \approx 2\alpha^2 n \log n.
\]

The theorem thus shows that the large clades are rare, and do not contribute to the asymptotic distribution; however, when they appear, the large clades give a large (actually negative) contribution to \( X_n \), and as a result, half the variance of \( X_n \) comes from the large clades. (When there is a large clade, there is less room for other clades, so \( X_n \) tends to be smaller than usually.)

For higher moments, the large clades play a similar, but even more extreme, role. Note that (for \( n \geq 2 \)) with probability \( 2/n \), the root of \( T_n \) has outdegree 1, and then it is the unique maximal clade, and thus \( X_n = 1 \). Since \( \mathbb{E}X_n = \alpha n + O(1) \) by (1), we thus have \( X_n - \mathbb{E}X_n = -\alpha n + O(1) \) with probability \( 2/n \), and this single exceptional event gives a contribution \( \sim (-1)^k 2\alpha^k n^{k-1} \) to \( \mathbb{E}(X_n - \mathbb{E}X_n)^k \), which explains a fraction \( (k - 2)/k \) of the moment (4); in particular, this explains why the moment is of order \( n^{k-1} \).

For proofs and further details, see [6].

**References**


