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 $\mathcal{T}_n$ (which corresponds to the Yule–Harding model of a random phylogenetic tree) and the number of maximal clades $X_n := F(\mathcal{T}_n)$ in it. We consider asymptotics as $n \to \infty$.

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

$$\mathbb{E}\, X_n = \mathbb{E}\, F(\mathcal{T}_n) = \alpha n + O(1), \tag{1}$$

where that the mean number of maximal clades $\mathbb{E}\, X_n \sim \alpha n$, where

$$\alpha = \frac{1-e^{-2}}{4}. \tag{2}$$

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

$$\mathbb{E}(X_n - \mathbb{E}\, X_n)^2 \sim 4\alpha^2 n \log n, \tag{3}$$

and for any fixed integer $k \geq 3$,

$$\mathbb{E}(X_n - \mathbb{E}\, X_n)^k \sim (-1)^k \frac{2k}{k-2} \alpha^k n^{k-1}. \tag{4}$$

As a consequence of (3)–(4), the limit distribution of $F(\mathcal{T}_n)$ (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:

$$\frac{X_n - \mathbb{E}\, X_n}{\sqrt{2\alpha^2 n \log n}} \xrightarrow{\mathrm{d}} N(0,1). \tag{5}$$

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 $\operatorname{Var}(X_n^N) \sim 2\alpha^2 n \log n$ and*

$$\frac{X_n^N - \mathbb{E}\, X_n^N}{\sqrt{\operatorname{Var} X_n^N}} \xrightarrow{\mathrm{d}} N(0,1). \tag{6}$$

Furthermore, $X_n - X_n^N = o_{\mathrm{p}}\bigl(\sqrt{\operatorname{Var} X_n^N}\bigr)$ and $\mathbb{E}\, X_n - \mathbb{E}\, X_n^N = o\bigl(\sqrt{\operatorname{Var} X_n^N}\bigr)$, so we may replace X_n^N by X_n in the numerator of (6)*. However,*

$$\operatorname{Var}(X_n - X_n^N) \sim \operatorname{Var}(X_n^N) \sim 2\alpha^2 n \log n. \tag{7}$$

The theorem thus shows that the large clades are rare, and do not contribute to the asymptotic distribution; however, when they appear, the larges 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 $\mathcal{T}_n$ 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

[1] Michael G. B. Blum and Olivier François, Minimal clade size and external branch length under the neutral coalescent. *Adv. in Appl. Probab.* **37** (2005), no. 3, 647–662.

[2] Huilan Chang and Michael Fuchs, Limit theorems for patterns in phylogenetic trees. *J. Math. Biol.* **60** (2010), no. 4, 481–512.

[3] Michael Drmota, Michael Fuchs and Yi-Wen Lee, Limit laws for the number of groups formed by social animals under the extra clustering model. (Extended abstract.) *Proceedings, 2014 Conference on Analysis of Algorithms, AofA '14 (Paris, 2014), DMTCS Proceedings*, 2014.

[4] Eric Durand, Michael G. B. Blum and Olivier François, Prediction of group patterns in social mammals based on a coalescent model. *J. Theoret. Biol.* **249** (2007), no. 2, 262–270.

[5] Eric Durand and Olivier François, Probabilistic analysis of a genealogical model of animal group patterns. *J. Math. Biol.* **60** (2010), no. 3, 451–468.

[6] Svante Janson, Maximal clades in random binary search trees. Preprint, 2014. `arXiv:1408.6337`