THE INFINITE VOLUME LIMIT OF FORD'S ALPHA MODEL

SIGURÐUR ÖRN STEFÁNSSON

November 11, 2009

Abstract. We prove the existence of a limit of the finite volume probability measures generated by tree growth rules in Ford's alpha model of phylogenetic trees. The limiting measure is shown to be concentrated on the set of trees consisting of exactly one infinite spine with finite, identically and independently distributed outgrowths.

1. Introduction

Graphs are used in many fields of science to describe relationships between individuals and to model actual physical objects. The former case includes social networks [2], phylogenetic trees [3, 13, 14], the world-wide web [1] and much more. The latter case includes discrete objects such as macromolecules [10] and branched polymers [2]. The graphs can also serve as discrete approximations to inherently continuous objects, an example of this being triangulation of manifolds in quantum gravity, see e.g. [4].

Random graphs are commonly used to describe real deterministic networks. Interactions and relations in the networks can be complicated but their characteristics are in some cases captured by random graph models, defined by simple rules which are motivated by the nature of the real network. The alpha model, introduced by D. Ford in [13], is an example of a random graph model, intended to describe phylogenetic trees. It is a one parameter model of randomly growing, rooted, planar, binary trees with the following growth rules. Start from a single rooted edge and from a tree on n leaves, select individual internal edges with probability weight α and individual leaves with probability weight $1-\alpha$ where $0 \leq \alpha \leq 1$. Graft a new leaf to a selected edge and thus generate a tree on $n+1$ leaves, see Fig. 1.

Ford proved that the model is Markovian self-similar which means informally that a subtree below an edge is distributed identically to the whole tree, a more precise definition will be given in the main section. He also showed that typical distances in the trees scale as n^{α} with the system size n. The Hausdorff dimension of a randomly growing tree is defined to be d_H given that typical distances scale as n^{1/d_H}. Thus, in the alpha model $d_H = 1/\alpha$.

In a recent paper [14] the continuum limit of the model has been established in the context of fragmentation processes [5]. A generalization to multinary trees is introduced in [7] in the alpha-gamma model where in addition to the growth rules of the alpha model, edges can be grafted onto vertices, increasing their degree. The alpha-gamma trees are shown to be Markovian self-similar and a continuum limit is established.

Our motivation to study the alpha model comes from the fact that it is a certain limiting case of a model of random trees which grow by vertex splitting, introduced in [8] where the relation is explained. In general the vertex splitting model does not share some of the technically convenient properties of the alpha model such as Markovian self-similarity, and it is more difficult to do exact calculations. The hope is that some of these properties might hold asymptotically for large trees and therefore a good understanding of the alpha model could be helpful.

The purpose of this paper is to establish convergence of the finite volume measures generated by the alpha model to a measure on infinite trees. For $0 < \alpha \leq 1$, the infinite measure is shown to be concentrated on the set of trees consisting of exactly one infinite path from the root to infinity (spine) with finite, identically and independently distributed outgrowths.

2. Convergence of the finite volume measures

We start with a few definitions before presenting the model. In this paper we only consider rooted, binary, planar trees. Rooted means that we mark a single vertex of degree 1, binary means that vertices are only allowed to have degree 1 or 3 and the planarity condition distinguishes between left and right branchings. The root and vertices of degree 3 will be referred to as internal vertices and vertices of degree 1, besides the root, will be referred to as leaves. Denote the set of trees on n leaves by T_n and denote the set of all finite or infinite trees by T.

The alpha model is defined by probability distributions $\pi_{\alpha,n}$ on T_n, for $n \geq 1$, constructed in the following recursive way. Assign probability one to the unique trees in T_1 and T_2. Given $\pi_{\alpha,n}$ for some $n \geq 2$, $\pi_{\alpha,n+1}$ is generated by first selecting a tree $\tau \in T_n$ according to $\pi_{\alpha,n}$. Next an individual edge (a, b) is selected from τ with probability $\alpha/(n-\alpha)$ if a and b are internal vertices and with probability $(1-\alpha)/(n-\alpha)$ if one is an internal vertex and the other a leaf. The edge (a, b) is removed from τ and two new vertices c and d are introduced along with the edges (a, c), (c, b) and (c, d). Equal probability is assigned to left and right branching of the new edge (c, d). One can think about this procedure as grafting a new edge to an existing edge in τ, see Fig. 1. The probability of a tree $\tau' \in T_{n+1}$ is thus given

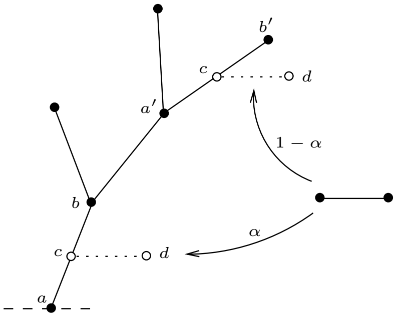

FIGURE 1. The grafting process. The link (a, b) is selected with probability weight α and the link (a', b') is selected with probability weight $1-\alpha$. The selected link is removed, two new vertices c and d and three new links are added as shown in the figure. In this example, a is the root which is indicated by a dashed line.

by

$$\pi_{\alpha,n+1}(\tau') = \sum_{\tau \in T_n} \pi_{\alpha,n}(\tau)\mathbb{P}(\tau \to \tau') \tag{1}$$

where $\mathbb{P}(\tau \to \tau')$ is the probability of growing the tree τ' from τ by the grafting process.

The model has a property called Markovian self-similarity [13] which is essential in the inductive proof of the theorem in this paper. Markovian self-similarity means that there exists a function $q_\alpha(\cdot,\cdot)$ such that for every finite tree τ_0 which branches at the nearest neighbour of the root to a left tree τ_1 and a right tree τ_2 (see Fig. 2) the following holds

$$\pi_{\alpha,|\tau_0|}(\tau_0) = q_\alpha(|\tau_1|,|\tau_2|)\pi_{\alpha,|\tau_1|}(\tau_1)\pi_{\alpha,|\tau_2|}(\tau_2) \tag{2}$$

where $|\tau|$ denotes the number of leaves in a tree τ. In words, this says that $q_\alpha(n_1,n_2)$ is the probability of a tree branching to subtrees of sizes n_1 and n_2. Furthermore, given that the subtrees are of these sizes they are distributed independently by π_{α,n_1} and π_{α,n_2}. The function q_α is explicitly known [13] and is given by

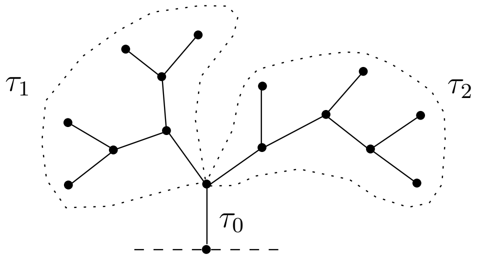

FIGURE 2. An example of a tree τ_0 which has a root indicated by the dashed line. The tree τ_0 branches at the nearest neighbour of the root to two subtrees, τ_1 to the left and τ_2 to the right as is indicated by the dotted lines.

$$q_\alpha(n_1,n_2) \quad = \quad \frac{n!\Gamma_\alpha(n_1)\Gamma_\alpha(n_2)}{n_1!n_2!\Gamma_\alpha(n)}\left(\frac{\alpha}{2}+\frac{(1-2\alpha)n_1n_2}{n(n-1)}\right)$$

where $n = n_1 + n_2$,

$$\Gamma_\alpha(n) = (n-1-\alpha)(n-2-\alpha)\cdots(2-\alpha)(1-\alpha), \quad \text{and} \quad \Gamma_\alpha(1) = 1. \tag{3}$$

Before proceeding to the theorem we give a short explanation of what is meant by convergence of probability measures. For a tree $\tau \in T$ let $B_R(\tau)$ be the subtree of τ which is spanned by the vertices at distance less than or equal to R from the root of τ. Define a metric d on T by

$$d(\tau,\tau') = \inf\left\{\frac{1}{1+R} \;\middle|\; B_R(\tau) = B_R(\tau')\right\}. \tag{4}$$

For some properties of the metric space (T,d) see [6, 11]. We will establish weak convergence, as $n \to \infty$ of the measures $\pi_{\alpha,n}$ viewed as probability measures on T, to a probability measure π_α. This means that for all bounded functions f which are continuous in the topology generated by the metric d

$$\int_T f(\tau) d\pi_{\alpha,n} \longrightarrow \int_T f(\tau) d\pi_\alpha, \qquad \text{as } n \longrightarrow \infty. \tag{5}$$

Theorem 2.1. *Let $0 < \alpha \leq 1$. The measures $\pi_{\alpha,n}$, viewed as probability measures on T, converge weakly, as $n \longrightarrow \infty$, to a probability measure π_α on infinite trees which is concentrated on the set of trees with one infinite rooted spine with finite outgrowths i.i.d. by*

$$\mu_\alpha(\tau) = \frac{\alpha \Gamma_\alpha(|\tau|)}{|\tau|!} \pi_{\alpha,|\tau|}(\tau). \tag{6}$$

The probabilities of right and left branching of outgrowths are equal (see Fig. 3).

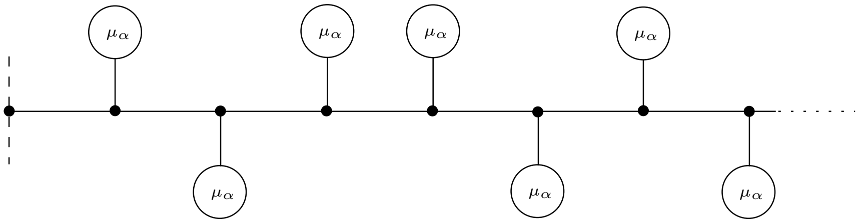

FIGURE 3. The infinite spine with finite μ_α–outgrowths.

Proof. We call the maximum graph distance from the root to a leaf in a tree, the height of the tree. Let $T^{(R)}$ be the set of rooted trees of height R. The metric space (T, d) is compact and therefore it is sufficient to show that for any $R \geq 1$ and any $\tau_0 \in T^{(R)}$ the sequence

$$\pi_{\alpha,n}(\{\tau | B_R(\tau) = \tau_0\}) =: \pi^{(R)}_{\alpha,n}(\tau_0) \tag{7}$$

converges to a limit $\pi^{(R)}_\alpha(\tau_0)$ as $n \longrightarrow \infty$ [11]. We show this by induction on R. For $R = 1$ this is trivial since $B_1(\tau) \in T^{(1)}$ for all τ. Now assume that for some R and all $\tau \in T^{(R)}$, $\pi^{(R)}_{\alpha,n}(\tau)$ converges as $n \longrightarrow \infty$. Choose a tree $\tau_0 \in T^{(R+1)}$ and without loss of generality, assume it branches at the nearest neighbour of the root to a left tree $\tau_1 \in T^{(R)}$ and a right tree $\tau_2 \in T^{(S)}$ (see Fig. 2) where $S \leq R$. Then

$$\begin{aligned} \pi^{(R+1)}_{\alpha,n}(\tau_0) &= \sum_{n_1+n_2=n} q_\alpha(n_1,n_2) \pi^{(R)}_{\alpha,n_1}(\tau_1) \pi^{(R)}_{\alpha,n_2}(\tau_2) \\ &= \frac{n!}{\Gamma_\alpha(n)} \Big(\frac{\alpha}{2} \sum_{n_1+n_2=n} \frac{\Gamma_\alpha(n_1)\Gamma_\alpha(n_2)}{n_1! n_2!} \pi^{(R)}_{\alpha,n_1}(\tau_1) \pi^{(R)}_{\alpha,n_2}(\tau_2) \\ &\quad + \frac{1-2\alpha}{n(n-1)} \sum_{n_1+n_2=n} \frac{\Gamma_\alpha(n_1)\Gamma_\alpha(n_2)}{(n_1-1)!(n_2-1)!} \pi^{(R)}_{\alpha,n_1}(\tau_1) \pi^{(R)}_{\alpha,n_2}(\tau_2) \Big). \end{aligned} \tag{8}$$

If $S < R$ then $\pi^{(R)}_{\alpha,n_2}(\tau_2) = 0$ when $n_2 > \ell(\tau_2)$ and it is obvious from the induction hypothesis that $\pi^{(R+1)}_{\alpha,n}(\tau_0)$ converges. Therefore assume that $S = R$.

Note that in (8) it always holds that either $n_1 \leq n-1$ and $n_2 \leq n$ or $n_2 \leq n-1$ and $n_1 \leq n$. Therefore we have the upper bound

$$\pi^{(R+1)}_{\alpha,n}(\tau_0) \le \frac{n!}{\Gamma_\alpha(n)} \sum_{n_1+n_2=n} \frac{\Gamma_\alpha(n_1)\Gamma_\alpha(n_2)}{n_1!n_2!}.$$

Terms in the sums in (8) for which $n_1 \ge \frac{n}{2}$ and $n_2 > A$ or $n_2 \ge \frac{n}{2}$ and $n_1 > A$ where $A > 1$ is some constant are therefore bounded from above by

$$\begin{aligned}\frac{2n!}{\Gamma_\alpha(n)} \sum_{\substack{n_1+n_2=n\\ n_1\ge n/2, n_2>A}} \frac{\Gamma_\alpha(n_1)\Gamma_\alpha(n_2)}{n_1!n_2!} &\le \frac{2n!\Gamma_\alpha([n/2])}{\Gamma_\alpha(n)[n/2]!} \sum_{n_2=A}^{\infty} \frac{\Gamma_\alpha(n_2)}{n_2!} \\ &\le C \sum_{n_2=A}^{\infty} \frac{\Gamma_\alpha(n_2)}{n_2!} \quad \underset{A\longrightarrow\infty}{\longrightarrow} \quad 0 \qquad (9)\end{aligned}$$

where C is a constant. The remaining contribution to (8) is from terms where $n_1 \ge \frac{n}{2}$ and $n_2 < A$ or $n_2 \ge \frac{n}{2}$ and $n_1 < A$. Notice that the second term in that contribution to (8) will be of one order lower in n than the first term. Therefore it is enough to show that the first term converges as $n \to \infty$ since then the second term clearly converges to zero. The contribution to the first term is

$$\begin{aligned}&\frac{n!}{\Gamma_\alpha(n)} \frac{\alpha}{2} \sum_{i=1}^{2} \sum_{\substack{n_1+n_2=n\\ n_j\le A, j\ne i}} \frac{\Gamma_\alpha(n_1)\Gamma_\alpha(n_2)}{n_1!n_2!} \pi^{(R)}_{\alpha,n_1}(\tau_1)\pi^{(R)}_{\alpha,n_2}(\tau_2) \\ &\underset{n\longrightarrow\infty}{\longrightarrow} \quad \frac{1}{2}\sum_{\substack{i=1\\ j\ne i}}^{2} \pi^{(R)}_{\alpha}(\tau_i) \sum_{m=1}^{A} \frac{\alpha\Gamma_\alpha(m)}{m!} \pi^{(R)}_{\alpha,m}(\tau_j) \\ &\underset{A\longrightarrow\infty}{\longrightarrow} \quad \frac{1}{2}\sum_{\substack{i=1\\ j\ne i}}^{2} \pi^{(R)}_{\alpha}(\tau_i) \sum_{m=1}^{\infty} \frac{\alpha\Gamma_\alpha(m)}{m!} \pi^{(R)}_{\alpha,m}(\tau_j). \qquad (10)\end{aligned}$$

In the first step we used the induction hypothesis. This is the limit of $\pi^{(R+1)}_{\alpha,n}(\tau_0)$ as $n \longrightarrow \infty$. The fact that the sum in (9) converges to zero as $A \to \infty$ proves that the measure is concentrated on the set of trees with exactly one infinite spine. The last sum in (10) shows that the distribution of the finite outgrowths is given by μ_α. □

3. Conclusions

We have shown that the finite volume measures $\pi_{\alpha,n}$ generated by the growth rules of Ford's alpha model converge, as $n \to \infty$, to a measure on infinite trees. The limiting measure is concentrated on the set of trees consisting of exactly one infinite spine with finite outgrowths, independently distributed by μ_α. The emergence of a single spine is well known from models of size conditioned critical Galton Watson trees [12]. The case $\alpha = 1/2$ is in fact a special case of such a tree. However, in the vertex splitting model it is possible that an infinite number of spines emerge. This happens for example in the special case of the preferential attachment model [9] and in the case $\alpha = 0$ in the alpha model. In both these cases the Hausdorff dimension is infinite. One interesting question is whether a finite Hausdorff dimension is equivalent to the emergence of a single spine and whether an infinite Hausdorff dimension is equivalent to the existence of infinite number of spines in the vertex splitting model.

An obvious next step is to use the formula for the limiting measure to calculate some global properties of the alpha trees such as the Hausdorff dimension and the spectral dimension. The Hausdorff dimension of an infinite random tree given by a probability distribution ν is defined as d_H if

$$\langle V_R \rangle_\nu \sim R^{d_H} \tag{11}$$

where $V_R(\tau)$ is the number of edges in a ball $B_R(\tau)$ and $\langle \cdot \rangle_\nu$ denotes expectation with respect to ν. The above definition should coincide with the one given by the scaling of a typical distance in a finite tree as discussed in the introduction. This will be checked explicitly in a forthcoming paper.

The spectral dimension of an infinite random tree as above is defined as d_s if

$$\langle p(t) \rangle_\nu \sim t^{-d_s/2} \tag{12}$$

where $p_\tau(t)$ is the probability that a simple random walk which starts at the root of a tree τ at time $t = 0$ is back at the root at time t. The techniques used in [12] give a way to estimate the spectral dimension of the alpha model from knowledge of the large R behaviour of the quantities $\langle |B_R| \rangle_{\mu_\alpha}$, $\mu_\alpha\{\tau | \text{ height of } \tau > R\}$ and $\langle |B_R|^{-1} \rangle_{\pi_\alpha}$. Using the formula for the limiting measure and the Markovian self-similarity properties of the outgrowths one can write recursion equations for generating functions of these quantities. Preliminary results indicate that indeed $d_H = 1/\alpha$ in agreement with the finite scaling definition and $d_s = 2/(1+\alpha)$. In the case $\alpha = 1$ this is trivially true and in the case $\alpha = 1/2$ the result is known to be true by connection to Galton Watson trees [12]. For other values of α this has not yet been proven.

Acknowledgment. This work is supported by the Eimskip Research Fund at the University of Iceland. I would like to thank Thordur Jonsson, François David and Mark Dukes for helpful discussions and valuable comments.

References

[1] Albert, R., Jeong, H. and Barabási, A. L., *Diameter of the world-wide web*, Nature, (1999). **401**, 130-131.

[2] Albert, R. and Barabási, A.-L., *Statistical mechanics of complex networks,* Rev. Mod. Phys. **74** (2002) 47.

[3] D. Aldous, *Probability distributions on cladograms. In Random Discrete Structures (Minneapolis, MN, 1993*, volume 76 of *Vol. Math. Appl.*, pages 1-18. Springer, New York, 1996.

[4] J. Ambjørn, B. Durhuus and T. Jonsson, *Quantum geometry: a statistical field theory approach,* Cambridge University Press, Cambridge (1997).

[5] J. Bertoin, *Random Fragmentation and Coagulation Processes*, Cambridge University Press, 2002, MR2253162.

[6] P. Billingsley. *Convergence of probability measures*, John Wiley and Sons, 1968.

[7] B. Chen, D. Ford and M. Winkel, *A new family of Markov branching trees, the alpha-gamma model*, Preprint, arXiv:0807:0554 [math.PR].

[8] F. David, M. Dukes, T. Jonsson and S. Ö. Stefánsson, *Random tree growth by vertex splitting*, J. Stat. Mech. (2009), no. 4, P04009.

[9] F. David, P. Di Francesco, E. Guitter and T. Jonsson, *Mass distribution exponents for growing trees*, J. Stat. Mech. (2007) P02011.

[10] F. David, C. Hagendorf and K. J. Wiese, *A growth model for RNA secondary structures*, J. Stat. Phys. 128 (2007) P02011.

[11] B. Durhuus. *Probabilistic aspects of infinite trees and surfaces*, Acta Physica Polonica B 34 (Oct. 2003) 4795.

[12] B. Durhuus, T. Jonsson and J. Wheater, *The spectral dimension of generic trees*, J. Stat. Phys. 128 (2007) 1237-1260.

[13] D. J. Ford, *Probabilities on cladograms: introduction to the alpha model*, Preprint, arXiv:math.PR/0511246.

[14] B. Hass, G. Miermont, J. Pitman and M. Winkel, *Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models*, Annals of Probability, 36(5), 1790-1837, 2008.

SCIENCE INSTITUTE, UNIVERSITY OF ICELAND, DUNHAGA 3, 107 REYKJAVÍK, ICELAND