

Structural Complexity of Random Binary Trees

J. C. Kieffer, E-H. Yang, and W. Szpankowski

Abstract—For each positive integer n , let T_n be a random rooted binary tree having finitely many vertices and exactly n leaves. We can view $H(T_n)$, the entropy of T_n , as a measure of the structural complexity of tree T_n in the sense that approximately $H(T_n)$ bits suffice to construct T_n . We are interested in determining conditions on the sequence $(T_n : n = 1, 2, \dots)$ under which $H(T_n)/n$ converges to a limit as $n \rightarrow \infty$. We exhibit some of our progress on the way to the solution of this problem.

I. INTRODUCTION

We define a *data structure* to consist of a finite graph (the structure part of the data structure) in which some of the vertices carry a label (these labels constitute the data part of the data structure). For example, in bioinformatics applications, the abstract notion of data structure can be used to model biological structures such as proteins and DNA strands. As another application example, in grammar-based data compression [2] [3], a data structure is used in which the graph is a directed acyclic graph from which the data string to be compressed is reconstructed by making use of the graph structure together with labels on the leaf vertices. Because of these and other potential applications of present day interest cited in [1], there is a need for developing an *information theory of data structures*.

Suppose one employs a random data structure model. Then each possible data structure would be generated with a certain probability according to the model. Let us denote the random data structure according to the probability model as (S, D) , where S denotes the structure part and D denotes the data part. We term the entropy $H(S)$ as the *structural complexity* of the random data structure (S, D) and the conditional entropy $H(D|S)$ as the *data complexity* of (S, D) . One challenge of the information theory of data structures would be to examine how the structural complexity and data complexity grow as the data structure increases in size without bound, and also to determine how these two complexity quantities compare to one another asymptotically. For example, in the grammar-based compression application, it is helpful if the ratio $H(S)/H(D|S)$ becomes negligibly small as the size of the data structure

J. C. Kieffer is with the Dept. of Electrical & Computer Engineering, University of Minnesota, Minneapolis, MN 55455.

E-H. Yang is with the Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, Ontario, CA N2L 3G1.

W. Szpankowski is with the Dept. of Computer Science, Purdue University, West Lafayette, IN 47907.

The work of the authors was supported in part by National Science Foundation Grants ECS-0830457, CCF-0513636, CCF-0503742, CCF-0830140, DMS-0800568, NIH Grant R01 GM068959-11, NSA Grant H98230-08-1-0092, AFOSR Grant FA8655-08-1-3018, and by Natural Sciences and Engineering Research Council of Canada Grants RGPIN203035-02 and RGPIN20305-06.

(S, D) grows [4].

Some initial advances have been made in developing the information theory of data structures. In the paper [1], the structural complexity of the Erdős-Rényi random graphs is examined. In the present paper, we examine the structural complexity of random binary trees. Letting T denote a random binary tree, its structural complexity is taken as its entropy $H(T)$. For a class of random binary tree models, we examine the growth of $H(T)$ as the number of edges of T is allowed to grow without bound. Letting $|T|$ denote the number of leaves of T , for some models we will be able to show that $H(T)/|T|$ converges as $|T| \rightarrow \infty$ (and identify the limit), whereas for some other models, we will see that $H(T)/|T|$ oscillates as $|T| \rightarrow \infty$.

Here is an outline of the rest of the paper. In Section II, we define the type of random binary tree probability model we will be using. In Section III, we specify the specific research questions we will be addressing regarding the asymptotic behavior of the entropy of a random binary tree. In Section IV, we give two formulas for computing the entropy of a random binary tree. In Sections V-VI, we address the Section III questions for some particular random tree models. In Section VII, our final section, we prove some bounds for the entropy of a random binary tree.

II. SPECIFICATION OF RANDOM BINARY TREE MODEL

Let \mathcal{T} be the set of all binary rooted trees having finitely many vertices. For each positive integer n , let \mathcal{T}_n be the subset of \mathcal{T} consisting of all trees in \mathcal{T} with n leaves. We define a random binary tree model to be a sequence $(T_n : n = 1, 2, \dots)$ in which each T_n is a random tree selected from \mathcal{T}_n . We shall be interested in a particular type of random binary tree model to be defined in this section.

If S is a finite set, we define a probability distribution p on S to be a mapping from S into the interval of real numbers $[0, 1]$ such that

$$\sum_{s \in S} p(s) = 1.$$

For each tree $t \in \mathcal{T}$, let $|t|$ denote the number of leaves of t . Furthermore, if t has at least two leaves, we define t^L (resp. t^R) to be the subtree of t rooted at the left (resp. right) child vertex of the root of t . Let T_n be any random tree selected from \mathcal{T}_n , where $n \geq 2$. Then T_n induces the probability distribution p_n on the set of positive integers $I_{n-1} = \{1, 2, \dots, n-1\}$ in which

$$p_n(i) = \text{Prob}[|T_n^L| = i], \quad i \in I_{n-1}. \quad (1)$$

In this way, each random binary tree model $(T_n : n =$

$1, 2, \dots$) induces the sequence of probability distributions $\{p_n\}_{n=2}^{\infty}$ in which each p_n satisfies (1).

Conversely, let $\{p_n\}_{n=2}^{\infty}$ be any sequence in which each p_n is a probability distribution on the set of positive integers I_{n-1} . There will be a multiplicity of random binary tree models $(T_n : n = 1, 2, \dots)$ which induce the sequence $\{p_n\}_{n=2}^{\infty}$ as described in the preceding paragraph. We will pick one of these random binary tree models and refer to it as *the binary random tree model* induced by $\{p_n\}_{n=2}^{\infty}$. We give both a probabilistic and a dynamic description of this binary random binary tree model.

A. Probabilistic Description

Let P be a probability distribution on \mathcal{T}_k (for fixed $k \geq 1$). Tree T selected randomly from \mathcal{T}_k is defined to have probability distribution P if

$$\text{Prob}[T = t] = P(t), \quad t \in \mathcal{T}_k.$$

Given a sequence $\{p_n\}_{n=2}^{\infty}$ in which each p_n is a probability distribution on I_{n-1} , the random binary tree model $(T_n : n = 1, 2, \dots)$ induced by $\{p_n\}$ is described probabilistically according to (a.1)-(a.2) below.

(a.1): Let t_1 be the unique tree belonging to \mathcal{T}_1 (t_1 consists of just one vertex, which is both root and leaf). Let P_1 be the probability distribution on \mathcal{T}_1 in which $P_1(t_1) = 1$. Then, random tree T_1 has probability distribution P_1 (that is, with probability one, random tree T_1 is equal to t_1).

(a.2): Let $n \geq 2$. Inductively, suppose that for each $k = 1, 2, \dots, n-1$, random tree T_k has probability distribution P_k on \mathcal{T}_k . Then, random tree T_n has the probability distribution P_n on \mathcal{T}_n in which

$$P_n(t) = p_n(|t^L|)P_{|t^L|}(t^L)P_{|t^R|}(t^R), \quad t \in \mathcal{T}_n. \quad (2)$$

B. Dynamic Description

Given a sequence $\{p_n\}_{n=2}^{\infty}$ in which each p_n is a probability distribution on I_{n-1} , let $(T_n : n = 1, 2, \dots)$ be the random binary tree model induced by $\{p_n\}$ as defined probabilistically in Section II.A. Random tree T_1 is of course equal to the unique tree in \mathcal{T}_1 with probability one. For each fixed $n \geq 2$, we can view random tree T_n as being dynamically generated as the result of several recursive steps, described in (b.1)-(b.3) below.

(b.1): In the first recursive step, the unique tree t_1 in \mathcal{T}_1 is labelled by assigning label n to the single vertex of t_1 . This labelled tree is the result of the first recursive step.

(b.2) : In each recursive step beyond the first step, one generates a labelled random tree by operating on the labelled tree T generated in the previous recursive step. Inductively, each vertex of T will have a positive integer label such that the label on each internal vertex is equal to the sum of the labels on its two children. Let v_1, v_2, \dots, v_k be the leaves of T , and let integers n_1, n_2, \dots, n_k be the corresponding leaf labels. For each leaf v_m for which label $n_m > 1$, randomly select integer i_m from the set I_{n_m-1} with probability $p_{n_m}(i_m)$ (independently of all other such leaves), and then grow two edges from leaf v_m with left edge terminating at a leaf of extended tree labelled i_m and

right edge terminating at a leaf of extended tree labelled $n_m - i_m$. The extended tree is the result of the current recursive step.

(b.3): Iterate the preceding recursive step. The recursion terminates with a binary tree having exactly n leaves, in which each leaf is labelled 1; this tree is T_n .

III. RESEARCH QUESTIONS

In this section, we state the research questions that we shall address concerning a random binary tree model. These questions all have to do with entropy, as defined in the usual information-theoretic sense as follows. If p is a probability distribution on a finite set S , then its entropy $h(p)$ is defined as

$$h(p) = \sum_{i \in S} -p(i) \log_2 p(i).$$

If X is a random object taking its values in a finite set S , then its entropy $H(X)$ is defined as

$$H(X) = \sum_{i \in S} -\text{Prob}[X = i] \log_2 \text{Prob}[X = i].$$

In both of the preceding entropy formulas, the quantity $0 \log_2 0$ is taken to be zero.

Let $(T_n : n = 1, 2, \dots)$ be a random binary tree model as defined in Section II. Then we define the *entropy rate* of the model to be the number

$$H_{\infty} = \limsup_{n \rightarrow \infty} H(T_n)/n.$$

The entropy rate H_{∞} tells us that we may describe the tree T_n for large n using no more than about H_{∞} bits per leaf (or $H_{\infty}/2$ bits per edge). There is in fact a natural arithmetic coding scheme which on average would generate a binary codeword for T_n of length about nH_{∞} for large n , and then tree T_n could be reconstructed from this codeword. (The recursive steps in building T_n in Section II.B are reflected as product factors in factorization of the probability of the realization of T_n ; this factorization is employed by the arithmetic coding scheme.)

We define the random binary tree model (T_n) to be *stable* if the sequence $\{H(T_n)/n\}$ has a limit as $n \rightarrow \infty$ (the limit of course must be H_{∞}). We define (T_n) to be *oscillatory* if the sequence $\{H(T_n)/n\}$ fails to have a limit as $n \rightarrow \infty$.

We are interesting in answering the following questions:

- Under what circumstances will a random binary tree model be stable or oscillatory?
- How can we compute the entropy rate of a random binary tree model?

These questions have not yet been completely answered. However, we have made some progress on these questions. Our progress is detailed in the subsequent sections of this paper.

IV. TWO FORMULAS FOR $H(T_n)$

Let $(T_n : n = 1, 2, \dots)$ be a random binary tree model induced by probability distributions $\{p_n\}_{n=2}^{\infty}$ as defined in

Section II. We present two formulas for computing the entropy $H(T_n)$ of tree T_n . The first formula is recursive, expressing $H(T_n)$ in terms of previous tree entropies. The second formula is direct, expressing tree entropies in terms of the entropies of the distributions $\{p_n\}$. These entropy formulas prove to be useful later on in examining asymptotic properties of the sequence $H(T_n)/n$ for various random binary tree models.

A. Recursive Formula

Let $(T_n : n = 1, 2, \dots)$ be a random binary tree model induced by probability distributions $\{p_n\}_{n=2}^\infty$. Starting with $H(T_1) = 0$, it is easy to see from formula (2) that the subsequent tree entropies are given by the recursion

$$H(T_n) = h(p_n) + \sum_{i=1}^{n-1} p_n(i)[H(T_i) + H(T_{n-i})], \quad n > 1.$$

Let p_n^* be the probability distribution on $I_{n-1} = \{1, 2, \dots, n-1\}$ in which

$$p_n^*(i) = (1/2)(p_n(i) + p_n(n-i)), \quad i \in I_{n-1}.$$

We can then rewrite our recursion more simply as

$$H(T_n) = h(p_n) + 2 \sum_{i=1}^{n-1} p_n^*(i)H(T_i), \quad n > 1. \quad (3)$$

B. Direct Formula

Definition. Let t be any tree in \mathcal{T} . For any integer $j \geq 2$, we define $N_j(t)$ to be the number of vertices of t for which the subtree of t rooted at the vertex has j leaves.

Theorem. Let $(T_n : n = 1, 2, \dots)$ be a random binary tree model induced by probability distributions $\{p_n\}_{n=2}^\infty$. Then

$$H(T_n) = \sum_{j=2}^n E[N_j(T_n)]h(p_j), \quad n \geq 1. \quad (4)$$

Proof. The proof is by induction. First, note that formula (4) is obviously true for $n = 1$. Let $n > 1$ be arbitrary, and, as our induction hypothesis, we assume formula (4) holds for all smaller values of n . We show that the formula holds for n as well. We denote random tree T_n more simply as T for the purposes of this proof. Let each vertex of T be labelled with positive integer label giving the number of leaves of the subtree of T rooted at that vertex. The root of tree T has two vertices as children, left vertex v_L labelled with random positive integer I_L and right vertex v_R labelled with random positive integer I_R . We have $I_L + I_R = n$ and p_n is the probability distribution of I_L . Recall our notation T^L (resp. T^R) for the subtree of T rooted at v_L (resp. v_R). We have

$$H(T) = H(I_L, T^L, T^R) = H(I_L) + H(T^L, T^R|I_L).$$

Since random trees T^L and T^R are conditionally independent given I_L ,

$$\begin{aligned} H(T^L, T^R|I_L) &= H(T^L|I_L) + H(T^R|I_L) \\ &= H(T^L|I_L) + H(T^R|I_R). \end{aligned}$$

Since we also have $H(I_L) = h(p_n)$, it follows that

$$H(T_n) = h(p_n) + H(T^L|I_L) + H(T^R|I_R). \quad (5)$$

I_L takes values in the set I_{n-1} ; let m be an arbitrary integer from this set. The conditional distribution of T^L given $I_L = m$ is the same as the distribution of T_m . Therefore, by the induction hypothesis,

$$\begin{aligned} H(T^L|I_L = m) &= H(T_m) = \sum_{j=2}^{\infty} E[N_j(T_m)]h(p_j) \\ &= \sum_{j=2}^{\infty} E[N_j(T^L)|I_L = m]h(p_j) \end{aligned}$$

It follows that

$$H(T^L|I_L) = \sum_{j=2}^{\infty} E[N_j(T^L)]h(p_j).$$

Similarly,

$$H(T^R|I_R) = \sum_{j=2}^{\infty} E[N_j(T^R)]h(p_j).$$

Note that

$$N_j(T^L) + N_j(T^R) = \begin{cases} N_j(T_n), & 1 \leq j \leq n-1 \\ 0, & j \geq n \end{cases}$$

Therefore,

$$\begin{aligned} H(T^L|I_L) + H(T^R|I_R) &= \sum_{j=2}^{\infty} E[N_j(T^L) + N_j(T^R)]h(p_j) \\ &= \sum_{j=2}^{n-1} E[N_j(T_n)]h(p_j). \end{aligned}$$

We conclude that

$$H(T_n) = \sum_{j=2}^n E[N_j(T_n)]h(p_j), \quad (6)$$

because

- the $j = n$ term on the right side of (6) is $h(p_n)$,
- the terms for which $j < n$ on the right side of (6) sum to $H(T^L|I_L) + H(T^R|I_R)$, and
- equation (5) holds.

The proof by induction is now complete.

V. TWO STABLE RANDOM TREE MODELS

In this section, we exhibit two stable random binary tree models, the *maximum entropy model* and the *binary search tree model*.

A. Maximum Entropy Model

Let K_n be the number of rooted binary trees with finitely many vertices and exactly n leaves. The *maximum entropy random tree model* is the model $(T_n : n = 1, 2, \dots)$ in

which the corresponding probability distributions $\{p_n\}_{n=2}^\infty$ are given by

$$p_n(i) = \frac{K_i K_{n-i}}{K_n}, \quad i \in I_{n-1}.$$

It is not difficult to show that T_n is equiprobable over the set of rooted binary trees with n leaves, that is,

$$\text{Prob}[T_n = t] = 1/K_n, \quad t \in \mathcal{T}_n.$$

It follows that

$$H(T_n) = \log_2 K_n, \quad n \geq 1.$$

It is well known [[5], Ch. 5] that

$$K_{n+1} = \frac{1}{n+1} \binom{2n}{n} = \frac{4^n}{\sqrt{\pi n^3}} \left(1 + O\left(\frac{1}{n}\right)\right). \quad (7)$$

It then follows that

$$\lim_{n \rightarrow \infty} H(T_n)/n = H_\infty = 2.$$

The maximum entropy model is therefore stable. It is called the maximum entropy model because 2 is the highest possible entropy rate of any of our random binary tree models.

B. Binary Search Tree Model

We consider the random binary tree model ($T_n : n = 1, 2, \dots$) in which the corresponding probability distributions $\{p_n\}_{n=2}^\infty$ are given by

$$p_n(i) = \frac{1}{n-1}, \quad i \in I_{n-1}.$$

This model is of interest because of its ubiquity in the computer science literature: if n keys are randomly permuted, then the binary search tree formed from them has the same distribution as the random tree T_n generated under this model (see Section 5.5 of [5]). For this reason, we call this model the binary search tree model.

Since $h(p_n) = \log_2(n-1)$ and $p_n^* = p_n$, the recursive formula (3) for $H(T_n)$ gives us

$$H(T_n) = \log_2(n-1) + \frac{2}{n-1} \sum_{i=1}^{n-1} H(T_i), \quad n > 1.$$

The unique solution to this recursion (starting with $H(T_1) = 0$) is

$$H(T_n) = \log(n-1) + 2n \sum_{k=1}^{n-2} \frac{\log k}{(k+1)(k+2)}, \quad n > 1.$$

We see from this that the binary search tree model is stable and

$$\lim_{n \rightarrow \infty} H(T_n)/n = H_\infty = 2 \sum_{k=1}^{\infty} \frac{\log_2 k}{(k+1)(k+2)} \approx 1.736.$$

VI. TWO OSCILLATORY RANDOM TREE MODELS

In this section, we exhibit two oscillatory random binary tree models, the *binomial model* and the *bisection model*.

A. Binomial Random Tree Model

The binomial random tree model ($T_n : n = 1, 2, \dots$) is the one in which the corresponding probability distributions $\{p_n\}_{n=2}^\infty$ are given by

$$p_n(i) = \binom{n-2}{i-1} (1/2)^{n-2}, \quad i \in I_{n-1}.$$

Theorem. For the binomial random tree model (T_n), there is a unique, nonconstant, continuous, periodic function P from the real line into itself, with period 1, such that

$$H(T_n)/n = P(\log_2 n) + o(n). \quad (8)$$

We have proved this theorem using the analytic depoissonization techniques of [6]. Moreover, we can identify the function P . The fact that P is nonconstant implies that the binomial random tree model is oscillatory.

Remark. Given one of our random binary tree models (T_n), we are currently investigating conditions under which there will exist a periodic function P satisfying (8). For such models, the question concerning whether the model is stable is easily answered: the model is stable if and only if the periodic function P is constant.

B. Bisection Random Tree Model

The bisection random tree model is important in connection with the bisection method of grammar-based data compression [3][4]. We define the bisection random tree model ($T_n : n = 1, 2, \dots$) (induced by probability distributions $\{p_n\}$) as follows. If integer $n > 1$ is odd, then

$$p_n(i) = \begin{cases} 1/2, & i = \lfloor n/2 \rfloor \\ 1/2, & i = \lceil n/2 \rceil \\ 0, & \text{elsewhere} \end{cases}$$

If integer $n > 1$ is even, then

$$p_n(i) = \begin{cases} 1, & i = n/2 \\ 0, & \text{elsewhere} \end{cases}$$

We then have the following recursion:

$$H(T_n) = \begin{cases} 0, & n = 1 \\ 1 + H(T_{\lfloor n/2 \rfloor}) + H(T_{\lceil n/2 \rceil}), & n > 1, \quad n \text{ odd} \\ 2H(T_{n/2}), & n > 1, \quad n \text{ even} \end{cases}$$

It is easy to see that

$$H(T_{2^k}) = 0, \quad k = 1, 2, 3, \dots, \quad (9)$$

and that

$$H(T_{3 \cdot 2^k}) = 2^k, \quad k = 1, 2, 3, \dots. \quad (10)$$

We easily conclude from (9)-(10) that the bisection random tree model is oscillatory. With some more work, omitted here, one can show that the bisection tree model has entropy rate $H_\infty = 1/2$.

VII. ENTROPY RATE BOUNDS

REFERENCES

In Section III, we posed the problem of determining a way to compute the entropy rate of a general random binary tree model. Although we are still trying to solve this problem, we have succeeded in obtaining the following result, which gives some bounds on entropy rate.

Theorem. Given the random binary tree model induced by the probability distributions $\{p_n\}_{n=2}^\infty$, let H_∞ be the entropy rate of the model. Then

(i): $H_\infty \leq \epsilon$ if $h(p_n) \leq \epsilon$ for all $n \geq 2$.

(ii): $H_\infty > 0$ if $\liminf_{n \rightarrow \infty} h(p_n) > 0$.

Remarks. Part(i) of the Theorem allows us to easily construct a variety of random binary tree models with arbitrarily small entropy rate by controlling the entropies of the p_n distributions which induce the model. Random binary tree models with small entropy rate prove to be useful in grammar-based data compression, as explained in [4]. Part(ii) of the Theorem tells us, for example, that the entropy rate of a random binary tree model will be positive whenever the entropies of the p_n distributions are strictly increasing.

Proof of Part(i). Let $(T_n : n = 1, 2, \dots)$ be a random binary tree model induced by $\{p_n\}_{n=2}^\infty$ in which every $h(p_n) \leq \epsilon$. Then by formula (4),

$$H(T_n) \leq \epsilon \sum_{j=2}^n E[N_j(T_n)] = \epsilon(n-1), \quad n \geq 2.$$

From this it is clear that

$$H_\infty = \limsup_{n \rightarrow \infty} H(T_n)/n \leq \epsilon.$$

Proof of Part(ii). Let $(T_n : n = 1, 2, \dots)$ be a random binary tree model induced by $\{p_n\}_{n=2}^\infty$ in which $\liminf_{n \rightarrow \infty} h(p_n) > 0$. There exists positive real number α and integer $\ell \geq 2$ such that

$$h(p_n) \geq \alpha, \quad \text{for all } n \geq \ell.$$

Applying formula (4),

$$H(T_n) = \sum_{j=2}^n E[N_j(T_n)]h(p_j) \geq \alpha \left[\sum_{j=\ell}^n E[N_j(T_n)] \right], \quad n \geq \ell.$$

It can be seen that there is a positive real number C such that

$$\sum_{j=\ell}^n N_j(t) \geq Cn, \quad n \geq \ell, \quad t \in T_n.$$

It follows that

$$H_\infty \geq \liminf_{n \rightarrow \infty} H(T_n)/n \geq \alpha C > 0.$$

- [1] Y. Choi and W. Szpankowski, "Compression of Graphical Structures," submitted to ISIT 2009.
- [2] J. C. Kieffer and E.-H. Yang, "Ergodic behavior of graph entropy," *Electron. Res. Announc. Amer. Math. Soc.*, Vol. 3, pp. 11–16, 1997.
- [3] J. C. Kieffer and E.-H. Yang, "Grammar-based codes: a new class of universal lossless source codes," *IEEE Trans. Inform. Theory*, Vol. 46, pp. 737–754, 2000.
- [4] J. C. Kieffer and E.-H. Yang, "Structured Grammar-Based Codes for Universal Lossless Data Compression," *Communications in Information and Systems*, vol. 2, pp. 29–52, 2002.
- [5] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*. Addison-Wesley Publishing Co., Inc., Reading, MA, 1996.
- [6] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, New York, 2001.