Journal of the Applied Mathematics, Statistics and Informatics (JAMSI), **2** (2006), No. 2

Expected number of vertices of random binary tree*

TOMÁŠ JURÍK

Abstract

Let B be the infinite complete random binary tree with the root r and p be the probability that two vertices of B are joined with a green-coloured edge. Let T be the connected subgraph (subtree) of B induced by the set of green edges which contains a vertex r. For some cases, graph T contains an infinite number of vertices. Throw these cases away. We prove that the average value of $|V(T)|$ is $(1-2p)^{-1}$.

Mathematics Subject Classification 2000: 05A15, 05C80, 11B37

Additional Key Words and Phrases: Binary trees, Catalan numbers, random graphs

1. INTRODUCTION

A binary tree is a rooted tree in which every node has at most two children. A complete binary tree is a special case of a binary tree, in which all the levels, except perhaps the last, are full. A perfect binary tree is a complete binary tree in which leaves (vertices with zero children) are at the same depth (distance from the root, also called height). Binary trees have an elegant recursive pointer structure, a useful tool to learn recursive pointer algorithms. Therefore they are the objects of wide studies. The binary search tree is one of the most frequently used structures in computer science. There exist many algorithms to generate such trees (see e.g. [3]). The reason to generate them effectively is Binary Search Tree, the most used technique to search some stored data. This search can be performed in $O(h)$ time, where h is the height of the tree.

Our approach to the binary trees is probabilistic. Problem solved in this paper came from [4]. Each edge of the infinite complete binary tree B with the root r is green-coloured with the probability p. We consider just a connected subtree T with green-coloured edges which contains the vertex r. Our question is: *How many vertices will such a random binary tree have?* Obviously, if $p = 1$ then T has an infinite number of vertices. We do not consider that cases, that trees T are thrown away from further computations. Using the main ideas from [1] we prove that the value of $|V(T)|$ is $(1-2p)^{-1}$ in Theorem 4.1.

2. RECURRENCE

We use direct approach to compute the expected number of vertices of the binary random tree. Let p_n be a probability that a constructed binary tree has exactly n

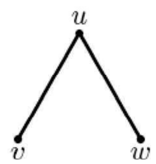
*Supported by VEGA grant No. 1/3022/06

vertices. Then the expected number (mean value) is

$$|V(T)| = E(p) = \sum_{n=1}^{\infty} n \cdot p_n.$$

We will find a recurrence relation for p_n. Obviously, $p_1 = (1-p)^2$. Let u be the root. Since $n \geq 2$ there exists at least one edge adjacent to u, there are two possibilities:

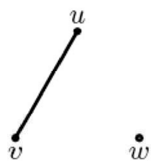
— At both adjacent vertices to u (each exists with probability p), named v and w there is a tree with k and $n-1-k$ vertices respectively.



This occurs with the probability

$$p^2\left(p_1 p_{n-2} + p_2 p_{n-3} + \cdots + p_{n-2} p_1\right).$$

— At just one vertex adjacent to u (with probability $p(1-p)$), named v or w there is a tree with $n-1$ vertices.



This has a probability

$$2p(1-p)p_{n-1}.$$

Summarizing we obtain the recurrence relation

$$p_n = 2p(1-p)p_{n-1} + p^2\left(p_1 p_{n-2} + p_2 p_{n-3} + \cdots + p_{n-2} p_1\right).$$

If we write some first terms down

$$p_2 = 2\,p(1-p)^3, \quad p_3 = 5\,p^2(1-p)^4, \quad p_4 = 14\,p^3(1-p)^5, \quad \ldots$$

we can easily prove by the induction that p_n has a form

$$p_n = c_n\, p^{n-1}(1-p)^{n+1}, \quad c_n \in \mathbb{R}.$$

Then our new goal is to solve the recurrence for terms c_n

$$c_n = 2c_{n-1} + \left(c_1 c_{n-2} + c_2 c_{n-3} + \cdots + c_{n-2} c_1\right),$$

Letting $c_0 = 1$ rewrite it into well-known recurrence formula

$$c_n = 2c_{n-1} + (c_1c_{n-2} + c_2c_{n-3} + \cdots + c_{n-2}c_1) = \sum_{k=0}^{n-1} c_kc_{n-1-k}, \quad n \geq 1. \tag{1}$$

We will operate with the following corollary which follows immediately from the definition of the generalized binomial coefficient.

Definition 2.1. The generalized binomial coefficient is defined as follows:

$$\binom{z}{k} = \prod_{n=1}^{k} \frac{z-k+n}{n} = \frac{z(z-1)(z-2)\cdots(z-k+1)}{k!}, \; z \in \mathbb{R}, \; k \in \mathbb{N}\cdot$$

LEMMA 2.2. *For all natural numbers n the equality is fulfilled*

$$\binom{2n}{n} = (-4)^n \binom{-\frac{1}{2}}{n}.$$

THEOREM 2.3 [2], PAGE 530. *The solution of the recurrence $c_0 = 1$ and $c_n = \sum_{k=0}^{n-1} c_kc_{n-1-k}$ for $n \geq 1$ are Catalan numbers*[1]

$$c_n = \frac{1}{n+1}\binom{2n}{n}.$$

3. THROWN AWAY TREES

In the previous section we proved that

$$p_n = p^{n-1}(1-p)^{n+1}\frac{1}{n+1}\binom{2n}{n}.$$

In this section we compute the probability that some random binary tree consists of infinite number of vertices. Let this probability be denoted as p_∞. Since the sum of all probabilities is one, we have

$$1 = p_\infty + \sum_{n=1}^{\infty} p_n = p_\infty + \sum_{n=1}^{\infty} p^{n-1}(1-p)^{n+1}\frac{1}{n+1}\binom{2n}{n}.$$

We want to compute the second summand. Consider the auxiliary real function $H(t)$ defined by

$$H(t) = \sum_{n=1}^{\infty} p^{n-1}(1-p)^{n+1}\frac{t^{n+1}}{n+1}\binom{2n}{n}.$$

One can see that $H(0) = 0$ and function $H(t)$ is well-defined for $p \in (0,1)$. We would like to compute $H(1)$ and we can do this using formula $H(1) = H(0) + \int_0^1 H'(t)dt$.

[1] http://mathworld.wolfram.com/CatalanNumber.html

Detailed computations (using Corollary 2.2) give us

$$\begin{aligned} H'(t) &= \sum_{n=1}^{\infty} p^{n-1}(1-p)^{n+1}t^n \binom{2n}{n} = \frac{1-p}{p}\sum_{n=1}^{\infty} p^n(1-p)^n t^n(-4)^n \binom{-\frac{1}{2}}{n} \\ &= \frac{1-p}{p}\sum_{n=1}^{\infty}(-4p(1-p)t)^n \binom{-\frac{1}{2}}{n} = \frac{1-p}{p}\left[(1-4p(1-p)t)^{-\frac{1}{2}}-1\right], \\ H(1) &= \frac{1-p}{p}\int_0^1 \left[\frac{1}{\sqrt{1-4p(1-p)t}}-1\right] dt \\ &= \frac{1-p}{p}\left[\frac{1}{2p(1-p)}(1-|1-2p|)-1\right] \\ &= -\frac{1}{2p^2}(|1-2p|-1)-\frac{1-p}{p}. \end{aligned}$$

If we consider two cases for an absolute value in the last term

$$|1-2p| = \begin{cases} 1-2p & 0 \le p \le 1/2 \\ 2p-1 & 1/2 < p \le 1 \end{cases}$$

we obtain the result in Theorem 3.1.

It is worth to note that the function $H'(t)$ is well-defined because the convergence of this sum is ensured by the condition $|4p(1-p)t| < 1$ it means $p \neq 1/2$. As we will see in Theorem 4.1 when $p \to 1/2$ then $H(1) \to \infty$.

THEOREM 3.1. *The expected average number of trees with infinite number of vertices is*

$$H(p) = \begin{cases} 1, & 0 \le p \le \frac{1}{2} \\ \left(\frac{1-p}{p}\right)^2, & \frac{1}{2} < p \le 1. \end{cases}$$

4. MAIN RESULT

This section is devoted to the proof of the main result stated in the following

THEOREM 4.1. *Let B be the infinite complete random binary tree with root vertex r and p be the probability that two vertices of B are joined with a green-coloured edge. Let T be the connected subgraph with all edges green which contains a vertex r. Then the average number of vertices of a graph T is*

$$|V(T)| = \frac{1}{|1-2p|} \cdot$$

We prove this result in two parts according to the value of p.

4.1 Case 1. $p < 1/2$

Immediately from the generalized binomial coefficient follows the following lemma.

LEMMA 4.2. *For each natural number n we have*

$$\binom{2n}{n-1} = \frac{2^n(-2)^{n-1}}{n+1}\binom{-\frac{3}{2}}{n-1}.$$

The first case is almost done. Theorem 3.1 states that for $p < \frac{1}{2}$ the probability that the tree has an infinite number of vertices is zero. So we need to compute $E(p) = \sum_{n=1}^{\infty} n \cdot p_n$.

$$\begin{aligned} E(p) &= \sum_{n=1}^{\infty} n \left[p^{n-1}(1-p)^{n+1} \frac{1}{n+1}\binom{2n}{n} \right] = \sum_{n=1}^{\infty} p^{n-1}(1-p)^{n+1}\binom{2n}{n-1} \\ &= \sum_{n=1}^{\infty} p^{n-1}(1-p)^{n+1} \frac{2^n(-2)^{n-1}}{n+1}\binom{-\frac{3}{2}}{n-1} \end{aligned}$$

We will use the same trick as in Section 3. Define a new real function F and compute $F(1)$

$$\begin{aligned} F(t) &= \sum_{n=1}^{\infty} p^{n-1}(1-p)^{n+1} \frac{2^n(-2)^{n-1}t^{n+1}}{n+1}\binom{-\frac{3}{2}}{n-1}, \\ F'(t) &= \sum_{n=1}^{\infty} p^{n-1}(1-p)^{n+1} 2^n(-2)^{n-1}t^n \binom{-\frac{3}{2}}{n-1} \\ &= 2t(1-p)^2 \sum_{n=1}^{\infty} (-4p(1-p)t)^{n-1} \binom{-\frac{3}{2}}{n-1} \\ &= 2t(1-p)^2(1-4p(1-p))^{-\frac{3}{2}}. \end{aligned}$$

$$\begin{aligned} F(1) &= F(0) + \int_0^1 F'(t)dt = 2(1-p)^2 \int_0^1 \frac{t\ dt}{(1-4p(1-p)t)^{\frac{3}{2}}} \\ &= \frac{2(1-p)^2}{4p^2(1-p)^2}\left(\frac{2p^2}{|1-2p|} + \frac{1-2p-|1-2p|}{|1-2p|} \right). \end{aligned}$$

If $p < \frac{1}{2}$ then the second summand in the braces is zero $(1 - 1p - |1-2p| = 0)$ therefore $F(1) = E(p) = (1-2p)^{-1}$.

4.2 Case 2. $p > 1/2$

As we see from Theorem 3.1 we have to throw some trees away. To fulfil the probability axioms ($\sum_{m=1}^{\infty} p'_n = 1$) we need to rescale the probabilities to obtain the new one p'_n

$$p'_n = \frac{p_n}{\sum_{m \geq 1} p_m} = \frac{p^{n-1}(1-p)^{n+1}\frac{1}{n+1}\binom{2n}{n}}{\left(\frac{1-p}{p}\right)^2} = p^{n+1}(1-p)^{n-1}\frac{1}{n+1}\binom{2n}{n}$$

which leads to

$$\sum_{n=1}^{\infty} n \cdot p'_n = E(1-p) = \frac{1}{1-2(1-p)} = \frac{1}{2p-1}, \quad \frac{1}{2} < p \leq 1.$$

The proof is completed.

ACKNOWLEDGMENTS

The author is grateful with an J. Plesník and E. Toman, both from Comenius University, for their helpful comments, corrections and remarks.

REFERENCES

[1] Palmer, E. M. : *Graphical evolution: an introduction to the theory of random graphs*, John Wiley & Sons, Inc., New York, NY, USA, 1985

[2] Gross J. L. and Yellen J. : *Handbook of Graph Theory*, CRC Press, Boca Raton, FL, USA, December 2003

[3] Mäkinen E. : *Generating Random Binary Trees - A Survey*, Information Sciences, April 1999, Volume 115, Issue 1-4, pp. 123-136.

[4] Shearer J. B. : *Ponder This*, http://www.research.ibm.com/ponder/ , Monthly Puzzles, November 2005

Tomáš Jurík,
Department of Mathematical Analysis and Numerical Mathematics,
Comenius University,
Mlynská dolina, 842 48 Bratislava, Slovak Republic
e-mail: jurik@fmph.uniba.sk

Received October 2006