Trees and Branching Processes

Brigitte Chauvin
Université de Versailles-St Quentin

January 5, 1998

[summary by Philippe Robert]

Abstract

A random tree is defined as an elementary event ω of a probability space $(\Omega, \mathcal{F}, P)$. The probability P depends on the random model of trees which is analyzed. The main results concerning the Galton-Watson processes are recalled. If for $n \in \mathbb{N}$, Z_n is the number of individuals of the N-th generation and m the average number of children generated by an individual, it is shown that the martingale (Z_n/m^n) plays an important role in the analysis of such processes.

The Catalan trees are seen as a particular case of Galton-Watson process. The height of a Catalan tree with n nodes is of the order $C\sqrt{n}$ (Flajolet-Odlyzko) and the number of external leaves has a limiting distribution (Kesten-Pittel).

The binary search trees are related to a branching random walk, hence to marked trees. The analysis of their height involves large deviations results for this random walk; for a binary search tree with n nodes, it is of the order $C \log n$ (Devroye, Biggins).

1. Probabilistic Model

Definition 1. If $Q = (q_i)$ is a probability distribution on $\mathbb{N}$ ($q_i \geq 0$ for $i \geq 0$ and $\sum_{i=0}^{+\infty} q_i = 1$), a Galton-Watson process with generating distribution Q is a sequence of random variables (Z_n) defined by

$$Z_0 = 1, \qquad Z_{n+1} = \sum_{i=1}^{Z_i} G_{in},$$

where the (G_{ij}), $i, j \in \mathbb{N}$ are independent identically distributed random variables with distribution Q.

For $n \in \mathbb{N}$, Z_n is the number of individuals at the n-th generation. By convention the generation 0 contains only the ancestor ($Z_0 = 1$) and the i-th individual of the n-th generation has G_{in} children.

The underlying tree structure of such a process is obvious. It is nevertheless convenient to reformulate these processes within the framework of trees [9]. A tree ω is a subset of

$$U = \{\emptyset\} \cup \bigcup_{n \geq 1} \mathbb{N}^{*n}$$

with the following properties:

1. $\emptyset \in \omega$, i.e. the ancestor is in the tree;
2. If $u \cdot v \in \omega$, then $u \in \omega$, ($u \cdot v$ denotes the concatenation of strings);

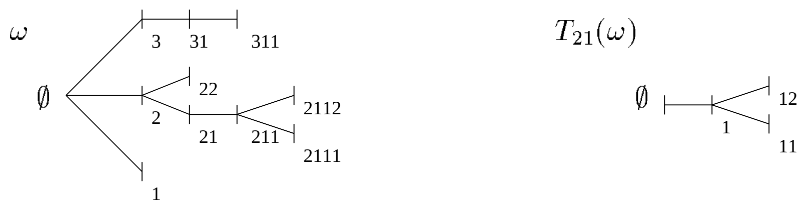

FIGURE 1. Trees as subsets of U

3. If $u \in \omega$ then there exists $N_u(\omega) \in \mathbb{N}$ such that $u \cdot j \in \omega$ if and only if $1 \leq j \leq N_u(\omega)$. The variable $N_u(\omega)$ is the number of children of the node u. By convention $N_\emptyset = N$.

With this notation, the tree of the Figure 1 can be represented as

$$\omega = \{\emptyset, 1, 2, 3, 21, 211, 2111, 2112, 22, 221, 31, 311\}.$$

If $u \in U$, $|u|$ will denote the length of the string u, in particular

$$H(\omega) = \sup\{|u|, u \in \omega\},$$

is the height of the tree ω and if $z_n(\omega) = \{u \in \omega, |u| = n\}$, then $Z_n(\omega) = \mathrm{Card}(z_n(\omega))$ is the number of individuals of generation n. Finally, if $u \in \omega$, $T_u(\omega)$ will denote the subtree containing the elements of ω whose prefix is u. In the example of Figure 1,

$$T_{21}(\omega) = \{\emptyset, 1, 11, 12\}.$$

Definition 2. A Galton-Watson tree with generating distribution Q is a probability distribution P on the set of trees such that

1. $P(N = k) = q_k$;
2. Conditionally on the event $\{N(\omega) = n\}$, the subtrees $T_1(\omega), T_2(\omega), \ldots, T_n(\omega)$ are independent with distribution P.

The first condition says that the number of children of the ancestor has distribution Q. The other condition gives an homogeneity property (the subtree $T_i(\omega)$ and ω have the same distribution for $i \leq n$). The independence of the behavior of the individuals, corresponds to the independence of the G_{i1}, $i = 1, \ldots, n$ in our first definition. From now on, (Z_n) denotes a Galton-Watson process associated to Q.

2. Limiting Behavior of Galton-Watson Trees

Notice that if $q_0 = P(N = 0) > 0$, then it is possible that an individual does not generate children at all. Consequently, a complete extinction of the family of the ancestor is also possible. The following proposition describes this phenomenon.

Proposition 1. *If $m = E(G_{11}) = \sum_{i=0}^{+\infty} iq_i$ is the average number of children per individual, then*

$$P\left(\sum_{n=0}^{+\infty} Z_n < +\infty\right) = q,$$

where q is the smallest solution $s \in [0,1]$ of the equation $\sum_{i=0}^{+\infty} q_i s^i = s$. If $m \leq 1$, the Galton-Watson becomes extinct with probability 1, that is, $q = 1$; and if $m > 1$ then $q < 1$.

We can now state a classical theorem for Galton-Watson processes.

Theorem 1. *The process $(W_n) = (Z_n/m^n)$ is a positive martingale with expected value 1, furthermore the sequence (W_n) is almost surely converging to a finite random variable W.*

Refinements. The following theorems give more insight on the behavior on the sequence (Z_n). There are three theorems, one for each of the three cases $m > 1$, $m = 1$ and $m < 1$.

Theorem 2 (Kesten-Stigum [7]). *If $m > 1$, the following conditions are equivalent*

1. *(Z_n/m^n) converge to W in $L_1(P)$;*
2. *$E(N \log N) = \sum_{i=2}^{+\infty} k \log(k) q_k < +\infty$;*
3. *$P(W = 0) = q$.*

The above result is mainly a strengthening of Theorem 1. It can be proved in an elegant way [8] with the formalism we described in the introduction. This proof uses a change of probability and the martingale (W_n).

The following theorem is more informative from a qualitative point of view. It says that in the critical case ($m = 1$) the variable Z_n grows linearly conditionally on $\{Z_n > 0\}$ (remember that in this case, almost surely $Z_n = 0$ for n large enough).

Theorem 3. *If $m = 1$ and $\sigma = \mathrm{Var}(N) < +\infty$, conditionally on the event $\{Z_n > 0\}$, the random variable Z_n/n converges in distribution to an exponential distribution with parameter $\sigma/2$.*

The same conditioning procedure as in the critical case does not lead to the same phenomenon in the sub-critical case ($m < 1$). Basically the conditioned variable Z_n stays bounded.

Theorem 4 (Yaglom [10]). *If $0 < m < 1$, then conditionally on the event $\{Z_n > 0\}$, the random variable converges in distribution to a finite random variable.*

3. Catalan Trees, Dyck Paths and Galton-Watson Processes

Definition 3.
1. A Catalan tree with n nodes is a random tree for the uniform distribution, that is, the probability of a tree ω is $P(\omega) = (n+1)/\binom{2n}{n}$, if $\mathrm{Card}(\omega) = n$ and 0 otherwise.
2. A Dyck path of length $2n$ is a positive path with the jumps 1, -1 starting at 0 and finishing at 0 for the $2n$-th jump.
3. An excursion of the simple random walk is the trajectory of the walk until it reaches 0 for the first time. A simple random walk is a walk which starts at 0 and whose jumps are 1 and -1 and equally likely.

Proposition 2.
- *The set of Catalan trees of size n and the set of Dyck paths with $2n$ steps have the same cardinality.*
- *The Galton-Watson process with $Q = (1/2^i)$ and the excursions of the simple random walk are isomorphic, i.e. there is a bijection which maps a Galton-Watson process to an excursion and preserves the distributions.*

Proof. The picture below shows how an excursion is transformed into a Galton-Watson process. □

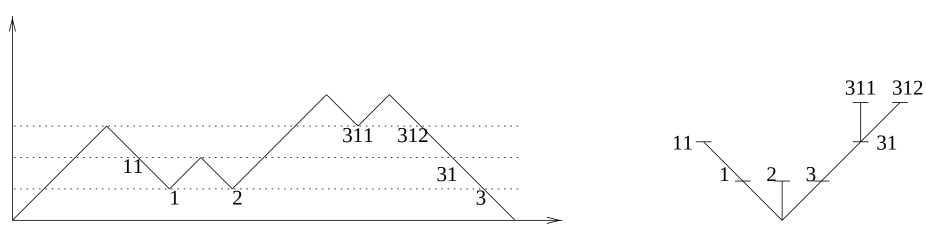

FIGURE 2. Equivalence between Galton-Watson processes and excursions

Remark. If one draws a contour starting at the left of the root of the tree in Figure 2 and following the vertices of the tree, when the contour arrives on the right of the root, its height will have performed the path followed by the random walk of Figure 2 above 1.

Bibliography

[1] Athreya (Krishna B.) and Ney (Peter E.). – *Branching processes.* – Springer-Verlag, New York, 1972, xi+287p. Die Grundlehren der mathematischen Wissenschaften, Band 196.

[2] Biggins (J. D.). – How fast does a general branching random walk spread? In *Classical and modern branching processes (Minneapolis, MN, 1994)*, pp. 19–39. – Springer, New York, 1997.

[3] Chauvin (Brigitte). – Sur la propriété de branchement. *Annales de l'Institut Henri Poincaré. Probabilités et Statistique*, vol. 22, n° 2, 1986, pp. 233–236.

[4] Devroye (Luc). – On the height of random m-ary search trees. *Random Structures & Algorithms*, vol. 1, n° 2, 1990, pp. 191–203.

[5] Flajolet (Philippe) and Odlyzko (Andrew). – The average height of binary trees and other simple trees. *Journal of Computer and System Sciences*, vol. 25, n° 2, 1982, pp. 171–213.

[6] Kesten (Harry) and Pittel (Boris). – A local limit theorem for the number of nodes, the height, and the number of final leaves in a critical branching process tree. *Random Structures & Algorithms*, vol. 8, n° 4, 1996, pp. 243–299.

[7] Kesten (Harry) and Stigum (B.). – A limit theorem for multidimensional Galton-Watson processes. *Annals of Mathematical Statistics*, vol. 37, 1966, pp. 1211–1223.

[8] Lyons (Russell), Pemantle (Robin), and Peres (Yuval). – Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *The Annals of Probability*, vol. 23, n° 3, 1995, pp. 1125–1138.

[9] Neveu (Jacques). – Arbres et processus de Galton-Watson. *Annales de l'Institut Henri Poincaré. Probabilités et Statistique*, vol. 22, n° 2, 1986, pp. 199–207.

[10] Yaglom (A. M.). – Certain limit theorems of the theory of branching random processes. *Doklady Akad. Nauk SSSR (N.S.)*, vol. 56, 1947, pp. 795–798.