# Distribution of tree parameters

Stephan Wagner

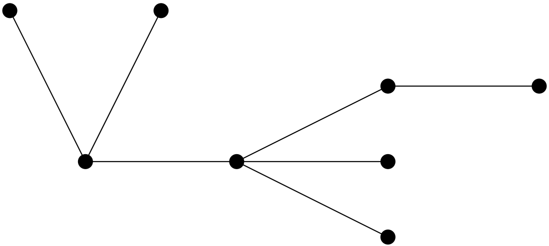
Stellenbosch University

Strathmore conference, 23 June 2017

# Trees

Trees are one of the simplest and best-studied classes of graphs. They are characterised by two main properties:

- They are connected.
- They do not have cycles.



Other properties follow, e.g. there is a unique path between any two vertices.

# Trees

Trees play a role in many applications, such as:

- data structures and algorithms,
- networks,
- phylogenetics,

and of course they are a very natural and interesting class of graphs to study from a purely mathematical point of view.

# Families of trees

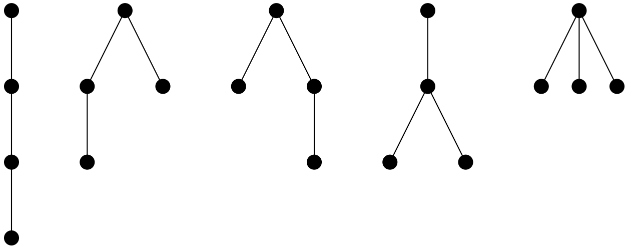
Trees can

- have labelled or unlabelled vertices,
- be rooted or unrooted,
- be plane or non-plane,
- have various restrictions (e.g. on vertex degrees).

Depending on these, many different classes of trees have been studied in the literature. The first question is typically *enumeration*.

# Families of trees

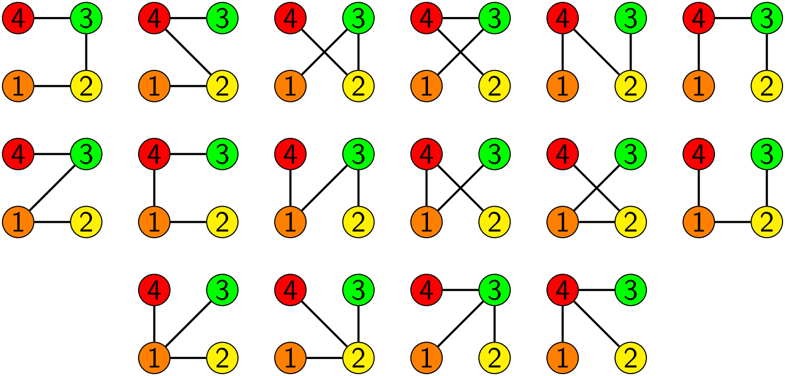
Plane trees: rooted trees embedded in the plane



The number of plane trees with $n$ vertices is the *Catalan number* $\frac{1}{n}\binom{2n-2}{n-1}$.

# Families of trees

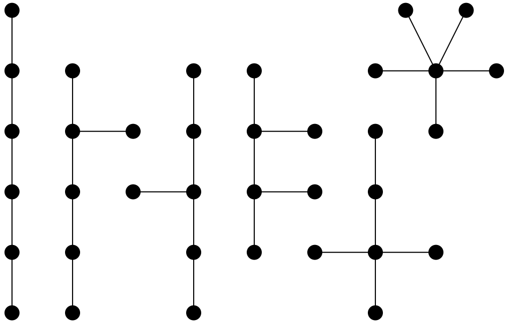
Labelled trees: each vertex has a unique label from $1$ up to $n$ (can be rooted or unrooted).



The number of labelled (unrooted) trees with $n$ vertices is $n^{n-2}$.

# Families of trees

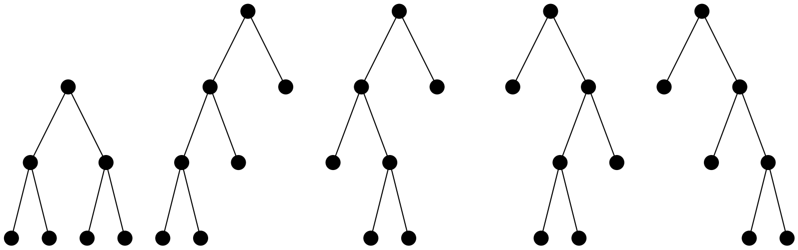
Unlabelled (unrooted) trees:



There is no simple formula for the number of unlabelled trees of a given size. The counting sequence starts $1, 1, 1, 2, 3, 6, 11, 23, 47, \ldots$, and there is an asymptotic formula for the number of trees with $n$ vertices: $0.53495 \cdot n^{-5/2} \cdot 2.95577^n$.

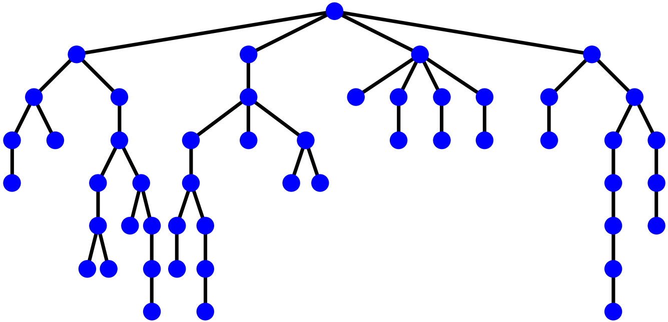
# Families of trees

Binary trees: rooted trees where every vertex is either a leaf or has exactly two children (left and right).



The number of binary trees with $n$ internal vertices is the *Catalan number* $\frac{1}{n+1}\binom{2n}{n}$.

# Random trees



A random tree with $50$ vertices.

# Random tree models

Random trees are studied in many areas, such as computational biology (phylogenetic trees) or the analysis of algorithms. Depending on the specific application, various random models have been brought forward, such as:

- Uniform models (e.g. uniformly random labelled or binary trees),
- Branching processes (e.g. Galton-Watson trees),
- Increasing tree models (e.g. recursive trees),
- Models based on random strings or permutations (e.g. tries, binary search trees).

# Uniform models

The simplest type of model uses the uniform distribution on the set of trees of a given order within a specified family (e.g. the family of all labelled trees, all unlabelled trees or all binary trees).

The analysis of such models typically involves exact counting and generating functions.

In particular, this is the case for *simply generated families of trees*.

# Simply generated families

On the set of all rooted ordered (plane) trees, we impose a weight function by first specifying a sequence $1 = w_0, w_1, w_2, \ldots$ and then setting

$$w(T) = \prod_{i \geq 0} w_i^{N_i(T)},$$

where $N_i(T)$ is the number of vertices of outdegree $i$ in $T$. Then we pick a tree of given order $n$ at random, with probabilities proportional to the weights. For instance,

- $w_0 = w_1 = w_2 = \cdots = 1$ generates random plane trees,
- $w_0 = w_2 = 1$ (and $w_i = 0$ otherwise) generates random binary trees,
- $w_i = \frac{1}{i!}$ generates random rooted labelled trees.

# Branching processes

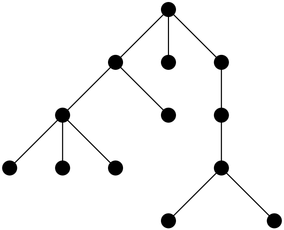
A classical branching model to generate random trees is the *Galton-Watson tree model*: fix a probability distribution on the set $\{0, 1, 2, \ldots\}$.

- Start with a single vertex, the root.
- At time $t$, all vertices at level $t$ (i.e., distance $t$ from the root) produce a number of children, independently at random according to the fixed distribution (some of the vertices might therefore not have children at all).
- A random Galton-Watson tree of order $n$ is obtained by conditioning the process.

Simply generated trees and Galton-Watson trees are essentially equivalent. For example, a geometric distribution for branching will result in a random plane tree, a Poisson distribution in a random rooted labelled tree.

# Simply generated and Galton-Watson trees

An example:



Consider the Galton-Watson process based on a geometric distribution with $P(X = k) = pq^k$ $(p = 1 - q)$.

The tree above has probability

$$p^7 (pq)^2 (pq^2)^2 (pq^3)^2 = p^{13} q^{12},$$

as does *every* tree with $13$ vertices.

# Random increasing trees

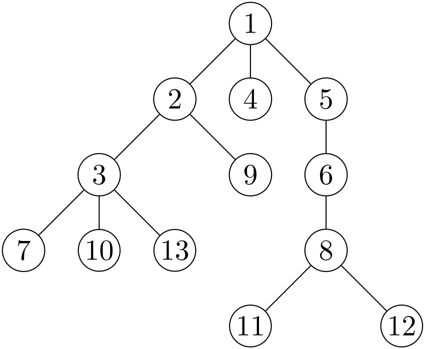
Another random model that produces very different shapes uses the following simple process, which generates *random recursive trees*:

- Start with the root, which is labelled 1.
- The $n$-th vertex is attached to one of the previous vertices, uniformly at random.

In this way, the labels along any path that starts at the root are increasing. Clearly, there are $(n-1)!$ possible recursive trees of order $n$, and there are indeed interesting connections to permutations. The model can be modified by not choosing a parent uniformly at random, but depending on the current outdegrees (to generate, for example, binary increasing trees).

# Random increasing trees

An example of an increasing tree:



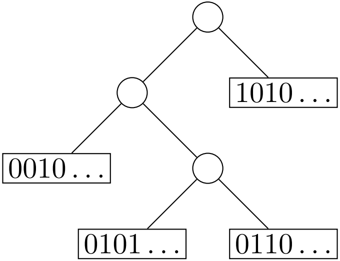
# Processes based on random strings

In computer science, tries (short for *retrieval trees*) are a popular data structure for storing strings over a finite alphabet. A random binary trie is obtained as follows:

- Create $n$ random binary strings of sufficient length, so that they are all distinct (for all practical purposes, one can assume that their length is infinite).
- All strings whose first bit is $0$ are stored in the left subtree, the others in the right subtree.
- This procedure is repeated recursively.

# Processes based on random strings

An example of a trie:



# Tree parameters

Many different parameters of trees have been studied in the literature, such as

- the number of leaves,
- the number of vertices of a given degree,
- the number of so-called fringe subtrees of a given shape,
- the height (maximum distance of a leaf from the root),
- the path length (total distance of all vertices from the root),
- the Wiener index (sum of distances between all pairs of vertices),
- the number of automorphisms,
- the total number of subtrees,
- the number of independent sets or matchings,
- ...

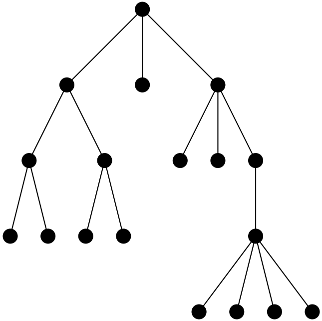
# A general question

Given a family of trees and a tree parameter, what can we say about . . .

- . . . the average value of the parameter among all trees with $n$ vertices?
- . . . the variance or higher moments?
- . . . the distribution?

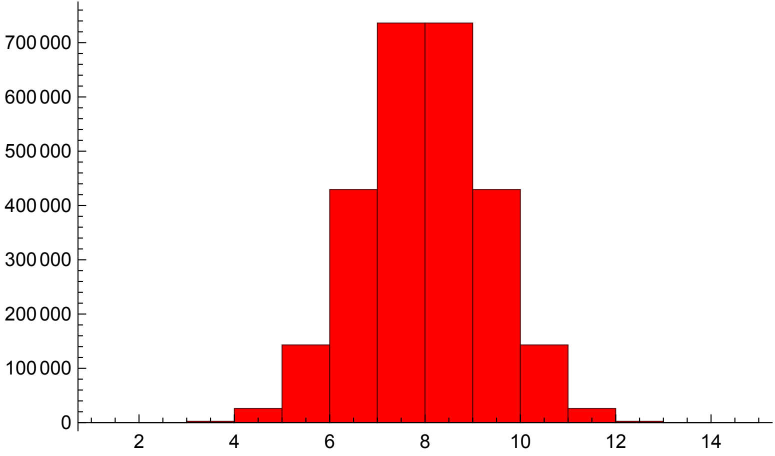
These questions become particularly relevant when $n$ is large.

# Some examples of parameters



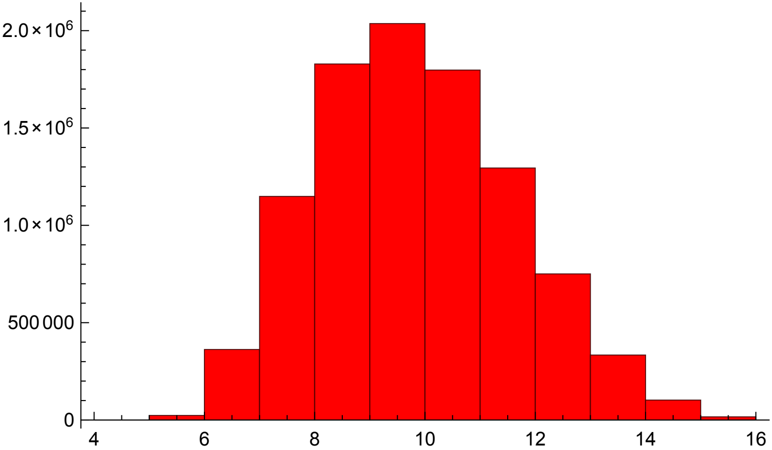
The tree above has $11$ leaves, $2$ “cherries”, height $4$, path length $44$, $384$ automorphisms and $3945$ subtrees.

# Distribution of parameters: some examples



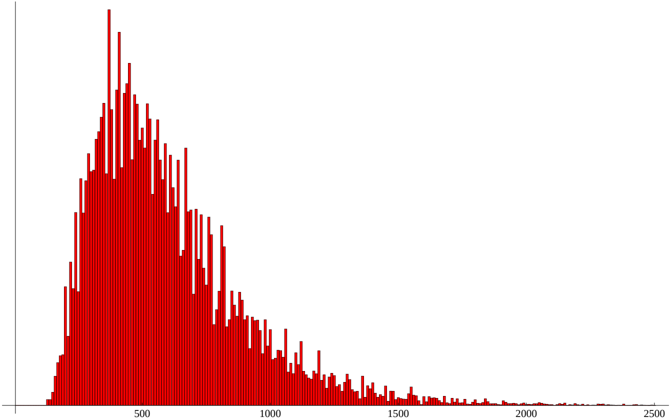

Distribution of the number of leaves among plane trees with $15$ vertices.

# Distribution of parameters: some examples




Distribution of the height among binary trees with $15$ internal vertices.

# Distribution of parameters: some examples




Distribution of the number of subtrees among labelled trees with $15$ vertices.

# Distributional results

In the following, let $\mathcal{F}$ be either a simply generated family of trees or the family of unlabelled trees (which is not simply generated, but has similar properties).

We consider a random element $T_n$ of $\mathcal{F}$ with $n$ vertices.

For some parameter $P$, what can we say about the distribution of $P(T_n)$?

# The number of leaves

Theorem (Drmota + Gittenberger 1999)

*For every family* $\mathcal{F}$*, there exist constants* $\mu > 0$ *and* $\sigma^2 > 0$ *such that the number of leaves* $L(T_n)$ *of a random tree* $T_n$ *in* $\mathcal{F}$ *has mean* $\mu_n \sim \mu n$ *and variance* $\sigma_n^2 \sim \sigma^2 n$*.*

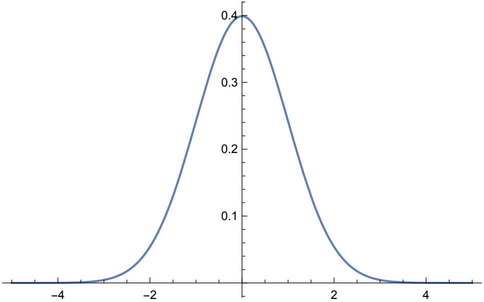
*Moreover, the renormalised random variable*

$$X_n = \frac{L(T_n) - \mu n}{\sqrt{\sigma^2 n}}$$

*converges weakly to a standard normal distribution* $N(0,1)$*.*

# The number of leaves

The theorem generalises to the number of vertices with a given degree or the number of fringe subtrees of a given shape.




The normal distribution: limiting distribution of the number of leaves.

# The height

Theorem (Flajolet, Gao, Odlyzko + Richmond 1993, Drmota + Gittenberger 2010)

*For every family $\mathcal{F}$, there exists a constant $\mu > 0$ such that the height $H(T_n)$ of a random tree $T_n$ in $\mathcal{F}$ has mean $\mu_n \sim \mu\sqrt{n}$.*

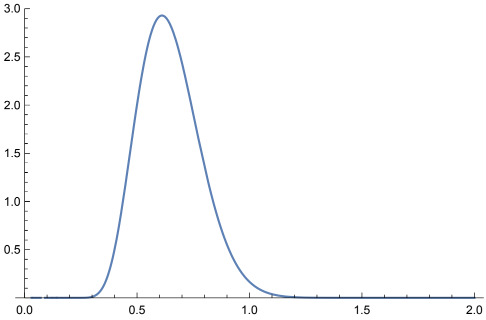
*Moreover, the renormalised random variable*

$$X_n = \frac{H(T_n)}{c\sqrt{n}},$$

*where $c = \frac{45\zeta(3)\mu}{2\pi^{5/2}}$, converges weakly to a so-called theta distribution, characterised by the density function*

$$f(t) = \frac{4\pi^{5/2}}{3\zeta(3)} t^4 \sum_{m \geq 1} (m\pi)^2 (2(m\pi t)^2 - 3) \exp(-(m\pi t)^2).$$

# The height




The theta distribution: limiting distribution of the height.

# Path length and Wiener index

Theorem (Takács 1993, Janson 2003, SW 2012)

*For every family $\mathcal{F}$, there exists a constant $\mu > 0$ such that the path length $D(T_n)$ and the Wiener index $W(T_n)$ of a random tree $T_n$ in $\mathcal{F}$ have means $\mu_n^D \sim \mu n^{3/2}$ and $\mu_n^W \sim \frac{\mu}{2} n^{5/2}$ respectively.*

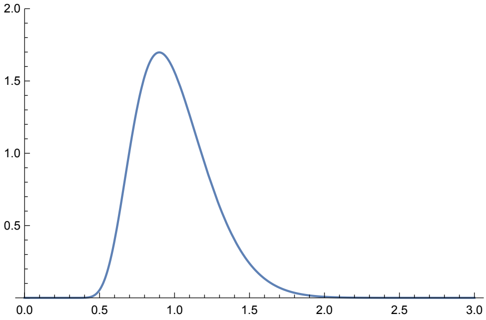
*Moreover, the renormalised random variables*

$$X_n = \frac{D(T_n)}{\mu n^{3/2}} \qquad \text{and} \qquad Y_n = \frac{W(T_n)}{\mu n^{5/2}}$$

*converge weakly to random variables given in terms of a normalised Brownian excursion $e(t)$ on $[0,1]$:*

$$\sqrt{\frac{8}{\pi}} \int_0^1 e(t)\,dt \quad \text{and} \quad \sqrt{\frac{8}{\pi}} \iint_{0<s<t<1} \big(e(s) + e(t) - 2 \min_{s \le u \le t} e(u)\big)\,ds\,dt.$$

# The path length




The Airy distribution: limiting distribution of the path length.

# The number of automorphisms

Theorem (Stufler + SW 2017+)

*For every family $\mathcal{F}$, there exist constants $\mu > 0$ and $\sigma^2 > 0$ such that the logarithm of the number of automorphisms $A(T_n)$ of a random tree $T_n$ in $\mathcal{F}$ has mean $\mu_n \sim \mu n$ and variance $\sigma_n^2 \sim \sigma^2 n$.*

*Moreover, the renormalised random variable*

$$X_n = \frac{\log A(T_n) - \mu n}{\sqrt{\sigma^2 n}}$$

*converges weakly to a standard normal distribution.*

# Additive tree parameters

## Definition

An invariant $F(T)$ defined for rooted trees $T$ is called *additive* if it satisfies the recursion

$$F(T) = \sum_{i=1}^{k} F(T_i) + f(T),$$

where $T_1, \ldots, T_k$ are the branches of the tree and $f(T)$ is a so-called toll function, which often only depends on the size of $T$.

## Remark

The recursion remains true for the tree $T = \bullet$ of order $1$ if we assume without loss of generality that $F(\bullet) = f(\bullet)$.

# Some examples

- The number of leaves, corresponding to the toll function

$$f(T) = \begin{cases} 1 & |T| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

- More generally, the number of occurrences of a fixed rooted tree $H$:

$$f(T) = \begin{cases} 1 & T = H, \\ 0 & \text{otherwise.} \end{cases}$$

- The number of vertices whose outdegree is a fixed number $k$:

$$f(T) = \begin{cases} 1 & \text{if the root of } T \text{ has outdegree } k, \\ 0 & \text{otherwise.} \end{cases}$$

# Some more examples

- The path length, i.e., the sum of the distances from the root to all vertices, can be obtained from the toll function $f(T) = |T| - 1$.
- The log-product of the subtree sizes, also called the "shape functional", corresponding to $f(T) = \log |T|$.

### Remark

Additive functionals also arise frequently in the study of divide-and-conquer algorithms, such as Quicksort.

# The number of subtrees

The number of subtrees is a more complicated example. It is somewhat more convenient to work with the number $s_1(T)$ of subtrees that contain the root (the difference turns out to be asymptotically negligible).

If $T_1, T_2, \ldots, T_k$ are the branches of $T$, then the following simple recursion holds:

$$s_1(T) = \prod_{i=1}^{k} (1 + s_1(T)).$$

Hence

$$\log(1 + s_1(T)) = \sum_{i=1}^{k} \log(1 + s_1(T_i)) + \log(1 + s_1(T)^{-1}).$$

This means that $\log(1 + s_1(T))$ is additive with toll function $f(T) = \log(1 + s_1(T)^{-1})$.

# A general result

Theorem (SW 2015, Janson 2016, Ralaivaosaona + Šileikis + SW 2017+)

*Under suitable technical conditions, an additive functional $F$ on a family $\mathcal{F}$ of trees satisfies a central limit theorem:*

*There exist constants $\mu$ and $\sigma^2$ such that mean and variance of $F(T_n)$ for a random tree $T_n$ in $\mathcal{F}$ are $\mu_n \sim \mu n$ and $\sigma_n^2 \sim \sigma^2 n$.*

*Moreover, the renormalised random variable*

$$X_n = \frac{F(T_n) - \mu n}{\sqrt{\sigma^2 n}}$$

*converges weakly to a standard normal distribution.*

# Examples covered

Many different examples are covered by one or more of the technical conditions, in particular:

- the number of leaves,
- the number of vertices of given degree,
- the number of fringe subtrees of a given type,
- the logarithm of the number of subtrees,
- the logarithm of the number of independent sets,
- the logarithm of the number of matchings.

# Conclusion

There are many different parameters of interest in connection with trees.

Their distributions lead to different types of limiting distribution as the size grows to infinity.

Several examples can be covered by general theorems on additive functionals.