CONCENTRATION PROPERTIES OF EXTREMAL PARAMETERS IN RANDOM DISCRETE STRUCTURES*

Michael Drmota

Inst. of Discrete Mathematics and Geometry

Vienna University of Technology, A 1040 Wien, Austria

michael.drmota@tuwien.ac.at

www.dmg.tuwien.ac.at/drmota/

* supported by the Austrian Science Foundation FWF, grant S9600.

Fourth Colloquium on Mathematics and Computer Science, Nancy, 18-22. 9. 2006

Outline of the Talk

- Types of Concentration
- Chromatic Number of Random Graphs
- Travelling Salesman Problem
- Longest Increasing Subsequence in Random Permutations
- Diameter and Maximum Degree in Random Graphs
- Height and Maximum Degree in Random Trees
- Height of Scale-Free Trees

Outline of the Talk

-
- Martingales $\longrightarrow$ concentration inequality
- Talagrand's *convex distance* $\longrightarrow$ concentration inequality
- Poisson transform, analytic methods
- Martingales, moment methods, ...
- Generating functions, analytic methods
- Generating functions, analytic methods

Conclusions

- Concentration (almost) always appears.
- Smaller extremal parameters are more concentrated than larger ones.
- Concentration is *easy* to prove
  (compared to the precise position of the mean).

Recent Books on Concentration

S. Boucheron, G. Lugosi, and O. Bousquet, *Concentration inequalities*, Lecture Notes in Computer Science 3176, Springer, Berlin, 2004.

D. P. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms*,
draft available at http://www.dsi.uniroma1.it/∼ale.

G. Lugosi, *Concentration-of-measure inequalities*,
draft available at http://www.econ.upf.es/∼lugosi.

C. McDiarmid, *Concentration*, Probabilistic methods for algorithmic discrete mathematics, 195–248, Algorithms Combin., 16, Springer, Berlin, 1998.

Types of Concentration

X_n ... non-negative random variable with $\boxed{\mathbf{E}\, X_n \to \infty}$

Concentration:

$$\boxed{\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{X_n}{a(n)} - 1\right| \geq \epsilon\right\} = 0}$$

for all $\epsilon > 0$ and some sequence $a(n)$ with $a(n) \to \infty$

Equivalently $X_n/a(n) \xrightarrow{\mathsf{d}} \delta_1$, usually $\boxed{a(n) = \mathbb{E}\, X_n}$.

Types of Concentration

X_n ... non-negative random variable with $\boxed{\mathbf{E}\, X_n \to \infty}$

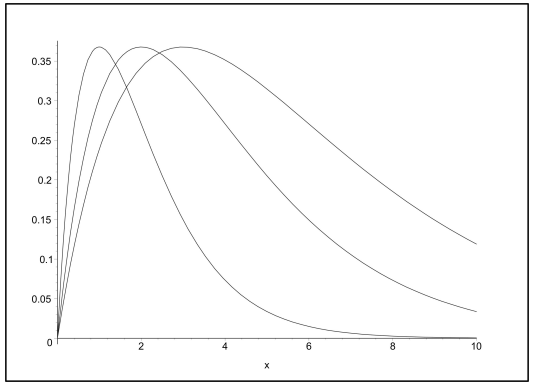
Type 1: No Concentration:

$$\boxed{\frac{X_n}{\mathbf{E}\, X_n} \not\xrightarrow{\mathsf{d}} \delta_1}$$

Typically:

$$\boxed{\frac{X_n}{\mathbf{E}\, X_n} \xrightarrow{\mathsf{d}} Y \ \ldots \text{ not concentrated at 1}}$$

and $\mathbf{E}\, X_n^2 \sim c \cdot (\mathbf{E}\, X_n)^2$ for some $c > 1$.



Types of Concentration

Type 2: Weak Concentration:

For all $\epsilon > 0$ there exists $K > 0$ such that

$$\boxed{\limsup_{n\to\infty} \mathbb{P}\left\{\left|\frac{X_n - a(n)}{b(n)}\right| \geq K\right\} \leq \epsilon.}$$

with $a(n) \to \infty$, $b(n) \to \infty$, and $b(n) = o(a(n))$

Usually one takes $a(n) = \mathbb{E}\, X_n$ and $b(n) = (\mathbb{V} X_n)^{1/2}$

If $\boxed{\mathbb{E}\, X_n^2 \sim (\mathbf{E}\, X_n)^2}$ the Chebyshev's inequality implies weak concentration.

Typically

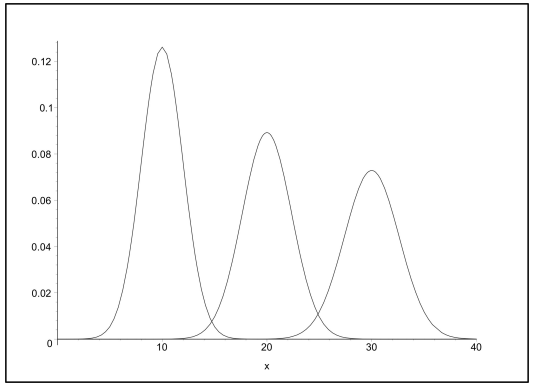
$$\frac{X_n - \mathbf{E}\, X_n}{\sqrt{\mathbf{Var}\, X_n}} \xrightarrow{\mathsf{d}} Y.$$

Types of Concentration

Type 2: Weak Concentration:

E.g. **Central Limit Theorem**

$$\frac{X_n - \mathbf{E}\, X_n}{\sqrt{\mathbf{Var}\, X_n}} \to N(0,1).$$



Types of Concentration

Type 3: Strong Concentration:

For all $\epsilon > 0$ there exists $K > 0$ with

$$\boxed{\limsup_{n \to \infty} \mathbb{P}\{|X_n - a(n)| \geq K\} \leq \epsilon}$$

for some sequence $a(n)$ with $a(n) \to \infty$.

Usually $a(n) = \mathbb{E}\, X_n$ or $a(n) =$ median of X_n and one has bounded centralized moments:

$$\boxed{\mathbb{E}\, |X_n - \mathbb{E}\, X_n|^d = O(1)} \qquad (d \geq 1).$$

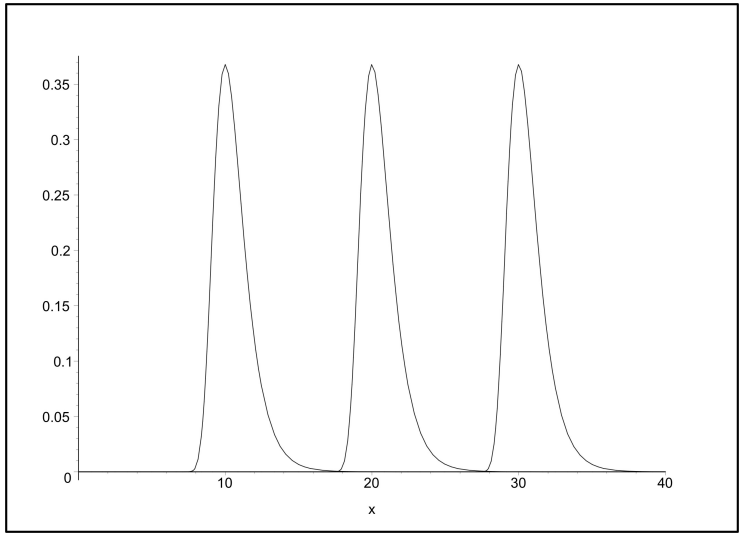
Types of Concentration

Type 3: Strong Concentration:

Typically: **travelling wave** $F(x)$

$$\boxed{\mathbb{P}\{X_n \leq k\} = F(k - m(n)) + o(1)}$$

($m(n)$ is close to the median of X_n)



Types of Concentration

Type 4: Very Strong Concentration:

Concentration on two (or finitely many values):

$$\boxed{\mathbb{P}\{m(n) \leq X_n \leq m(n) + L\} = 1 + o(1)}$$

with $m(n) \to \infty$ and some fixed L .

Chromatic Number of Random Graphs

Definition

Let n be a positive integer and p a real number with $0 \leq p \leq 1$.

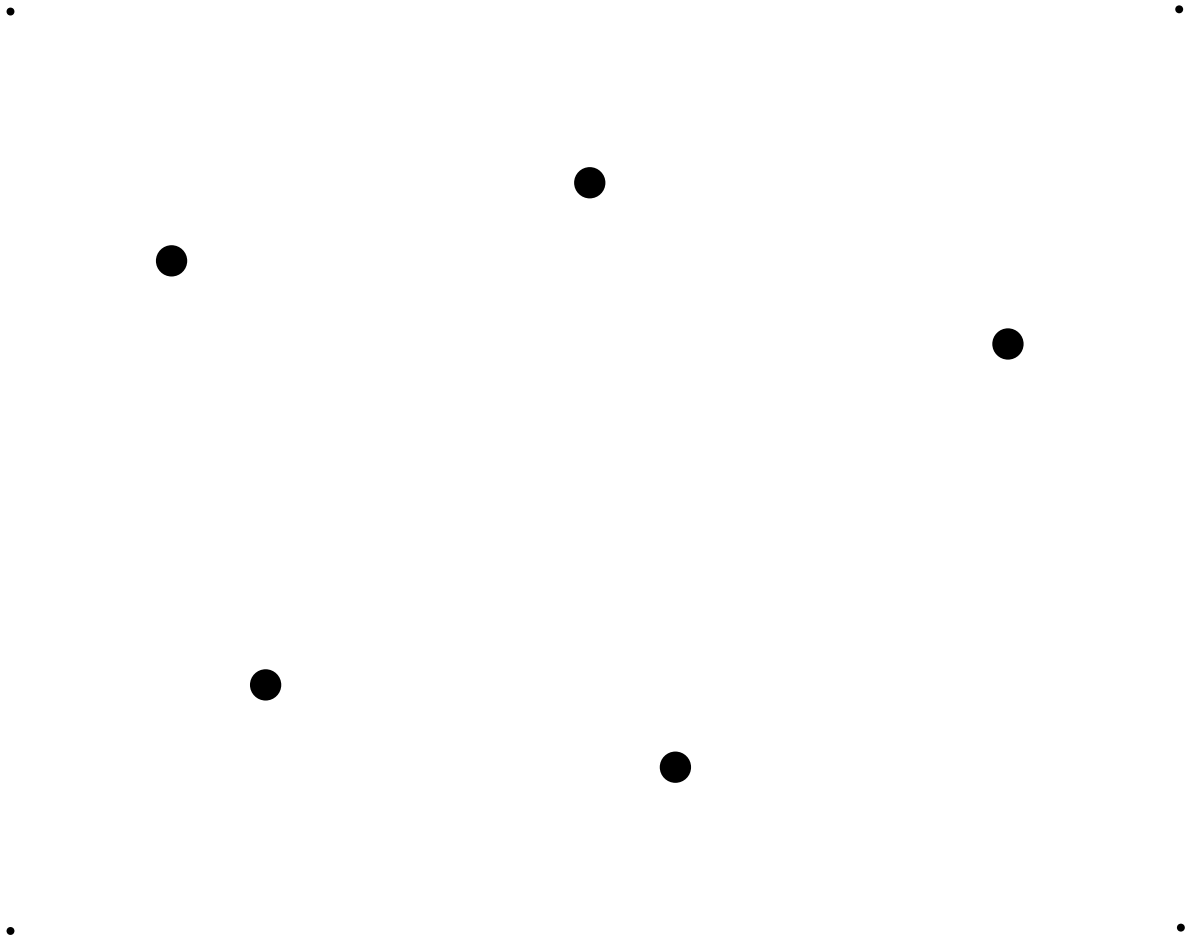
The **random graph** $G(n,p)$ *is a probability space over the set of graphs on the vertex set $\{1,2,\ldots,n\}$ determined by*

$$\mathbb{P}\{\{i,j\} \in G\} = p$$

for all possible (undirected) edges $\{i,j\}$ with $1 \leq i < j \leq n$ with these events mutually independent.

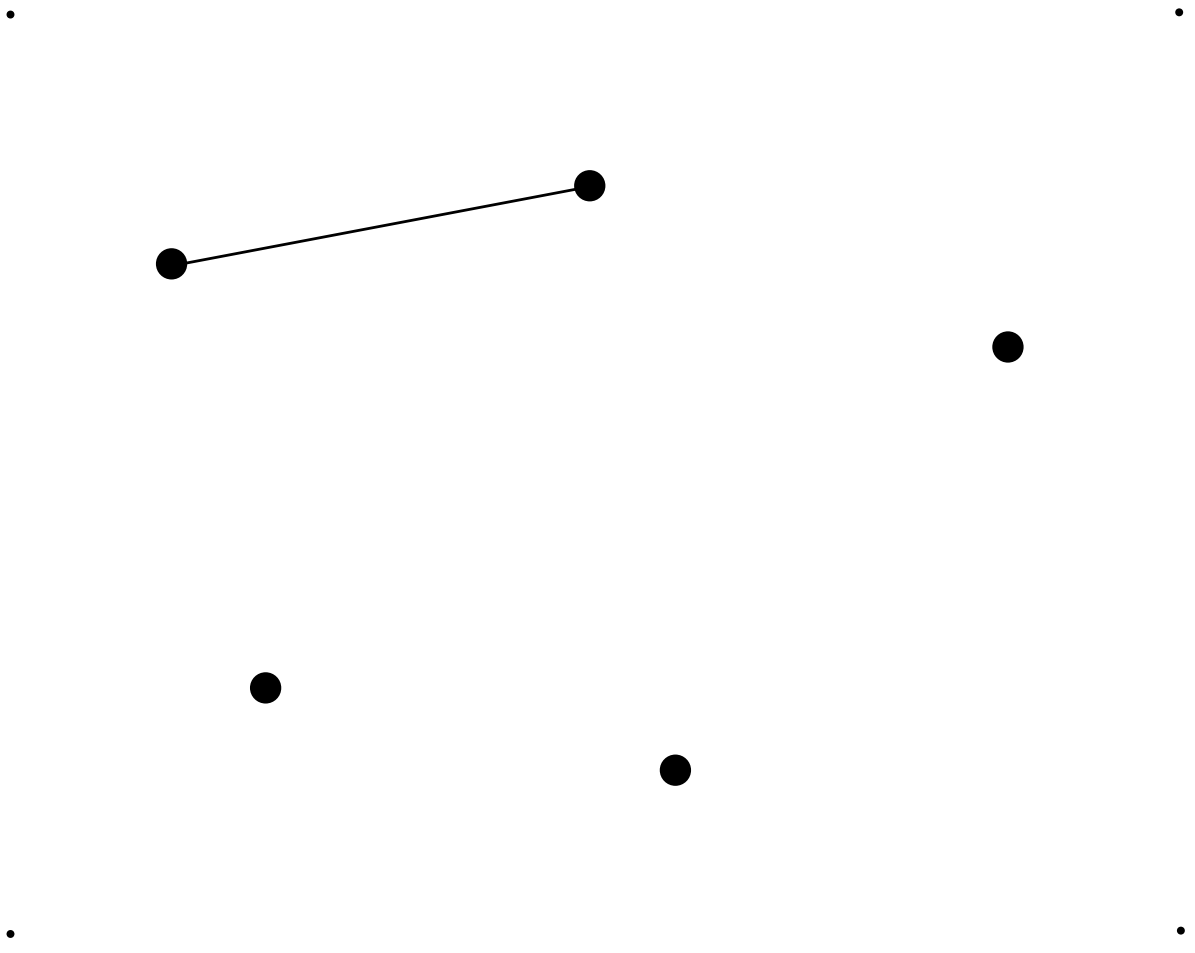
Chromatic Number of Random Graphs

Random Graph



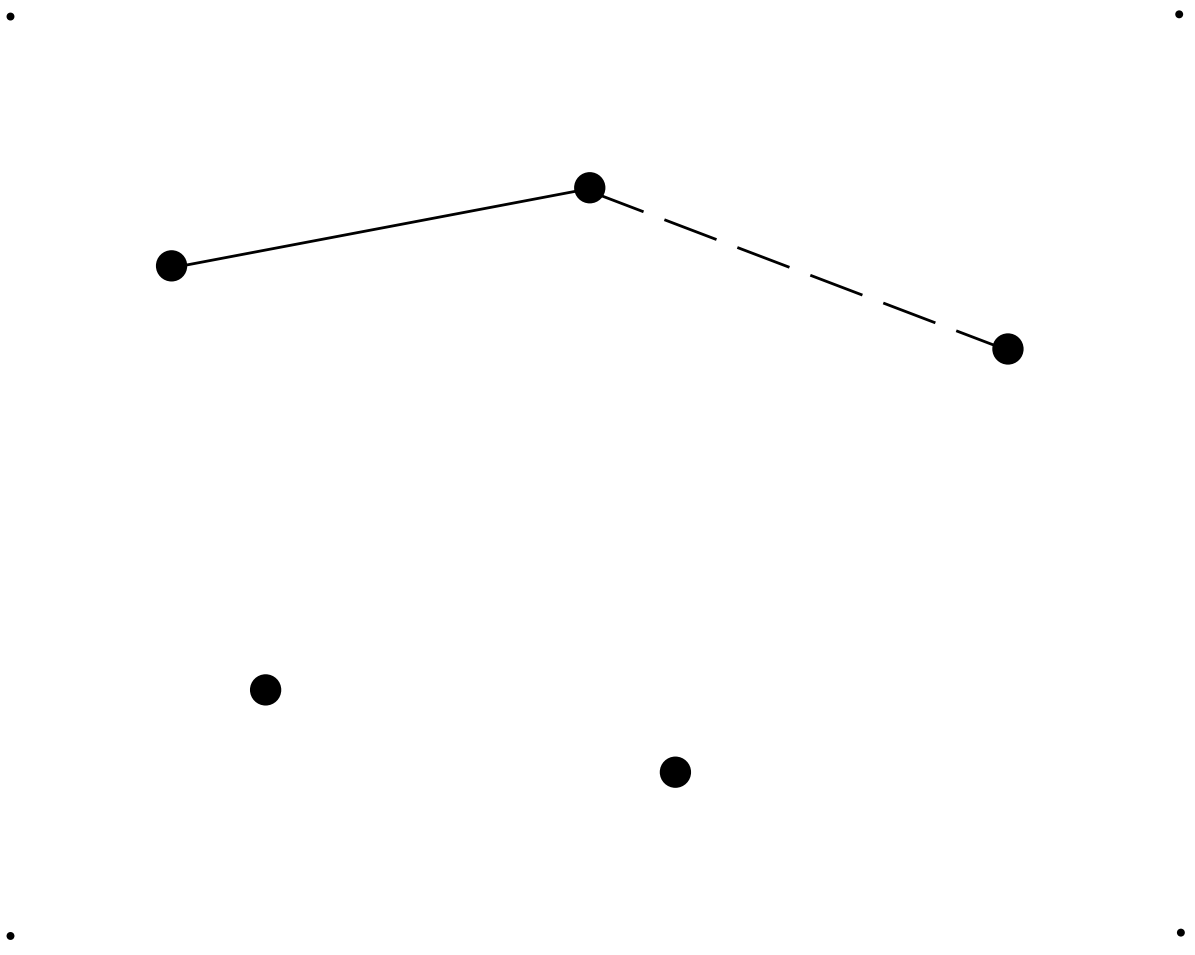
Chromatic Number of Random Graphs

Random Graph



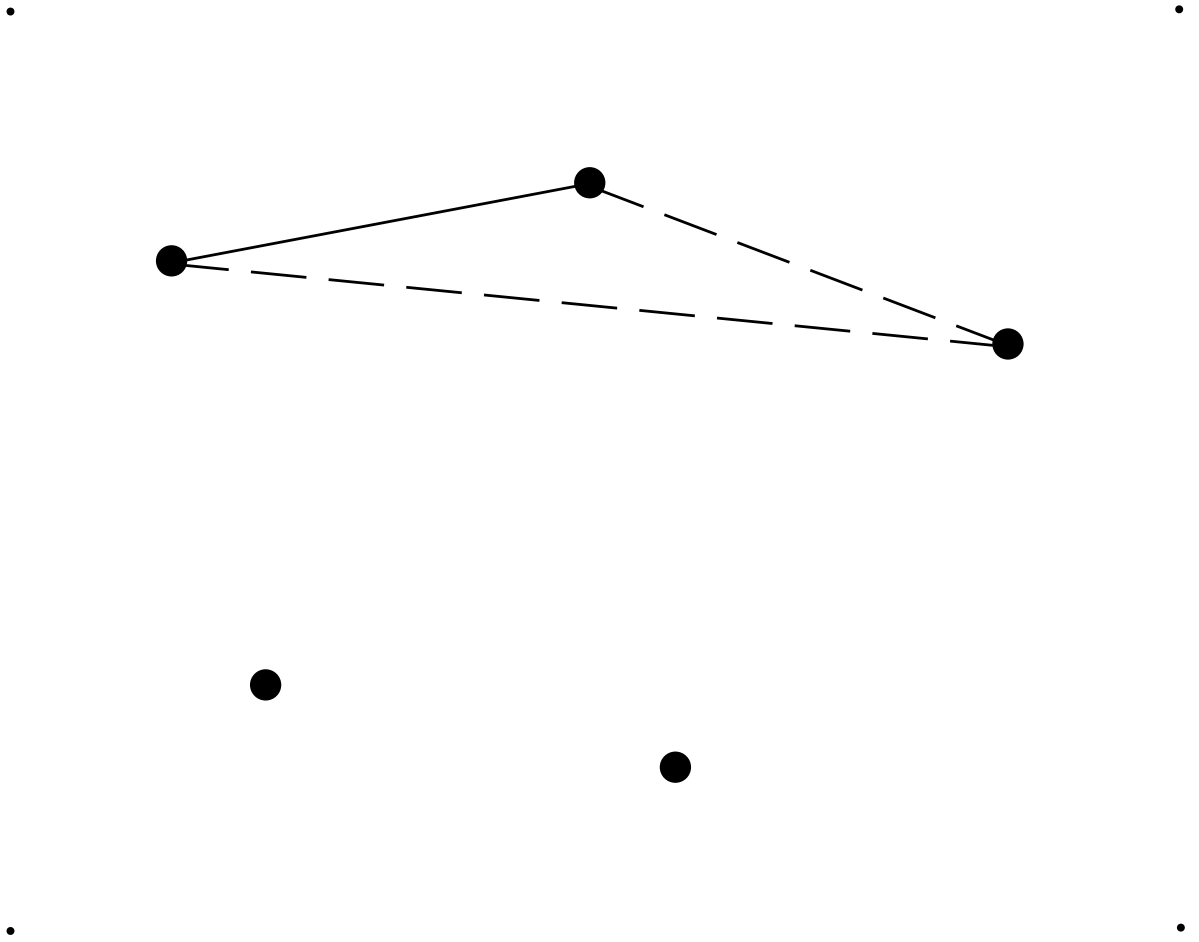
Chromatic Number of Random Graphs

Random Graph



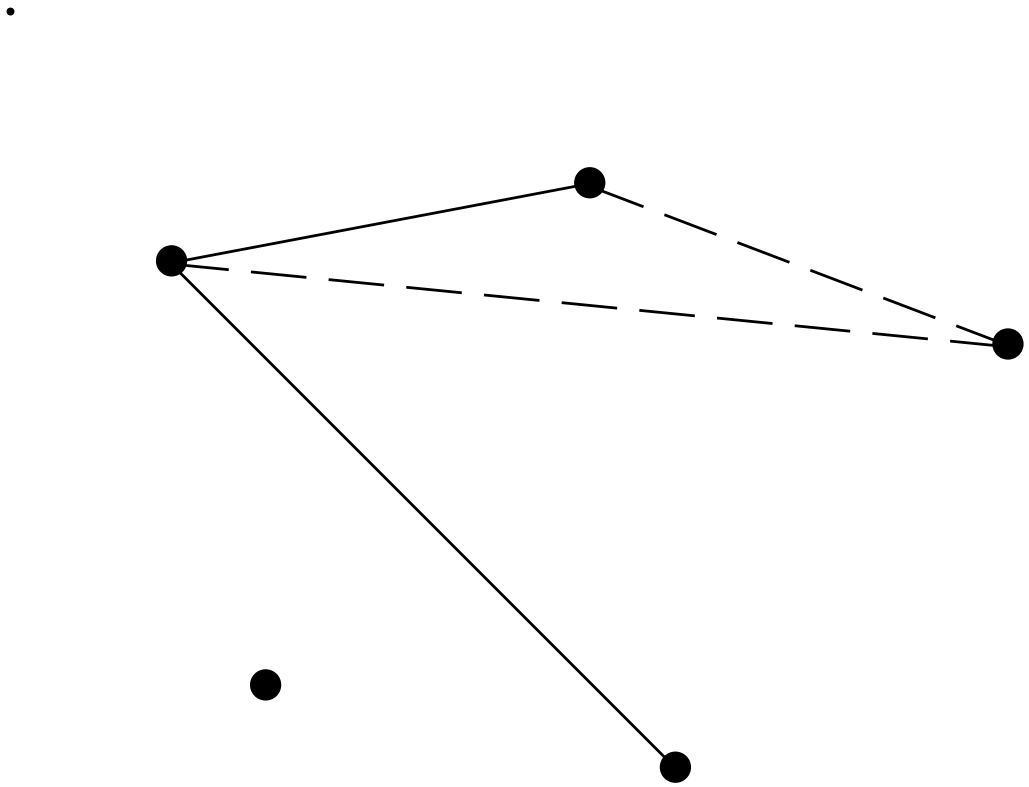
Chromatic Number of Random Graphs

Random Graph



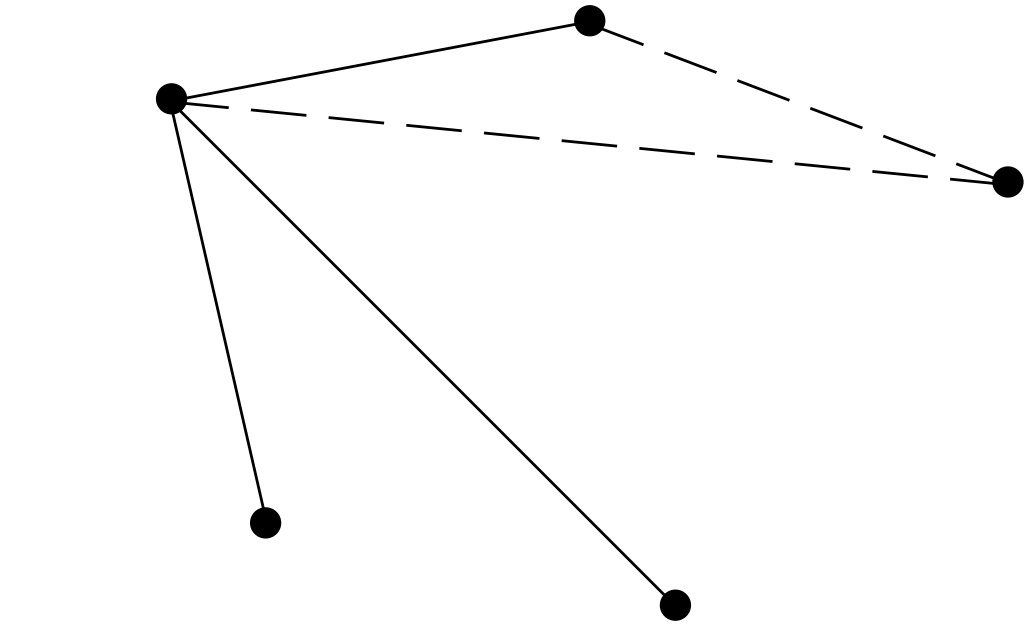
Chromatic Number of Random Graphs

Random Graph



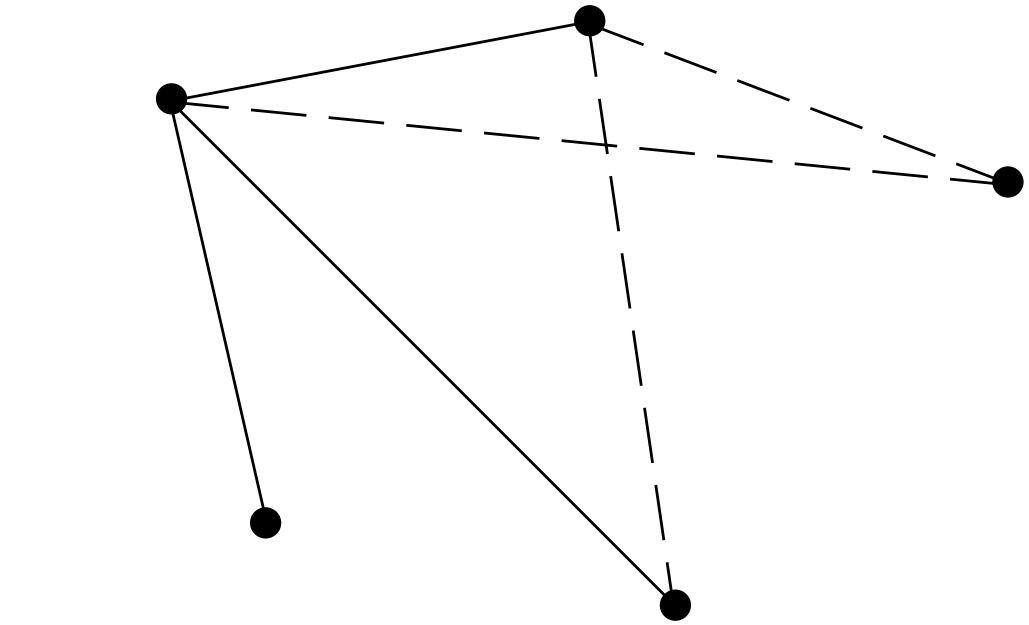
Chromatic Number of Random Graphs

Random Graph



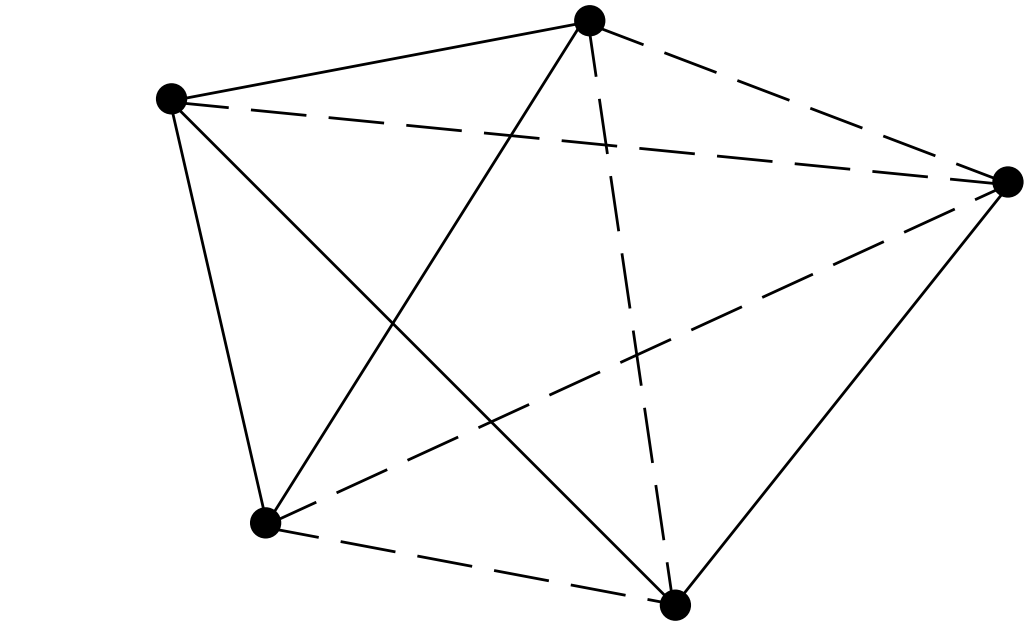
Chromatic Number of Random Graphs

Random Graph



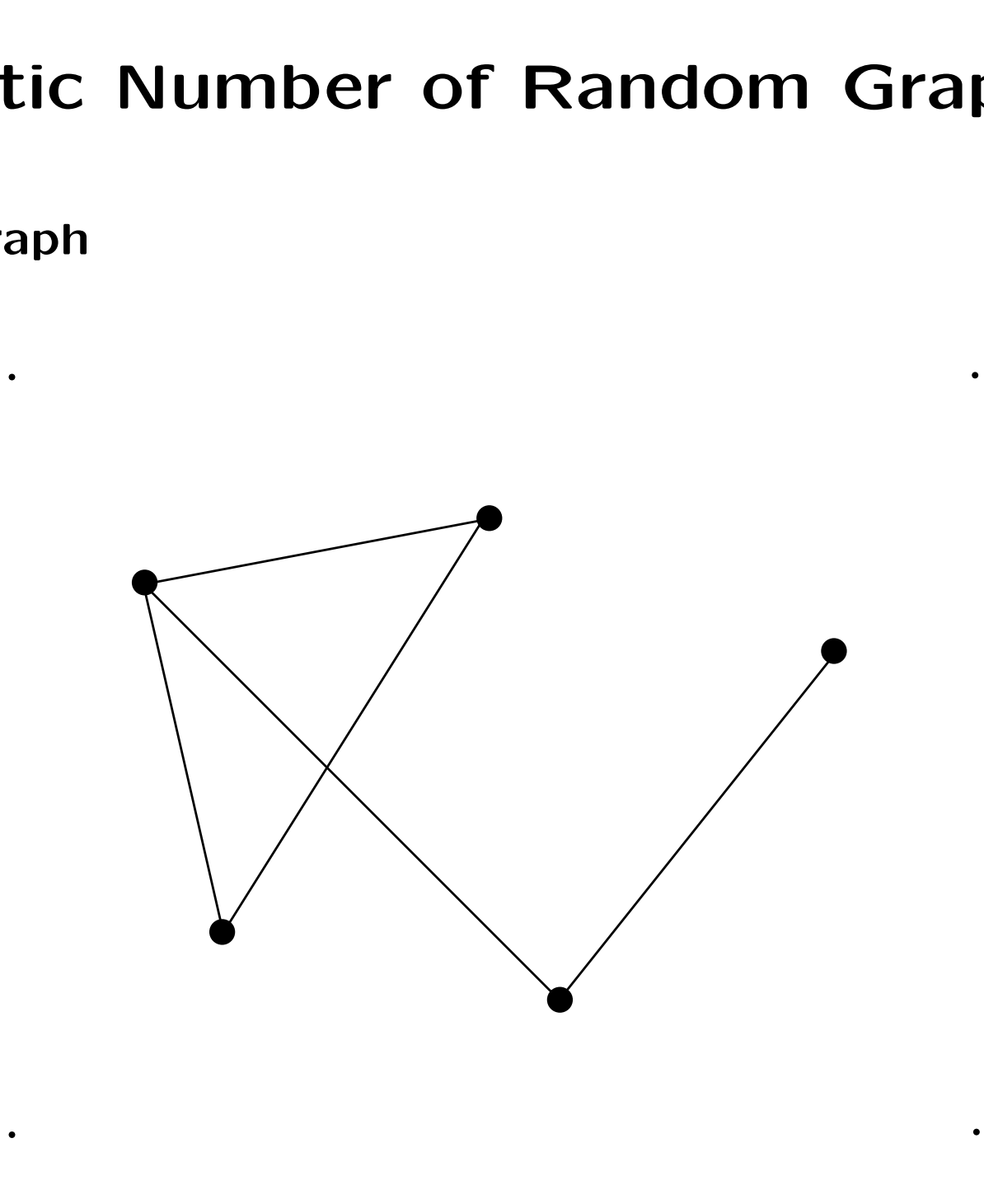
Chromatic Number of Random Graphs

Random Graph



Chromatic Number of Random Graphs

Random Graph



Chromatic Number of Random Graphs

Definition

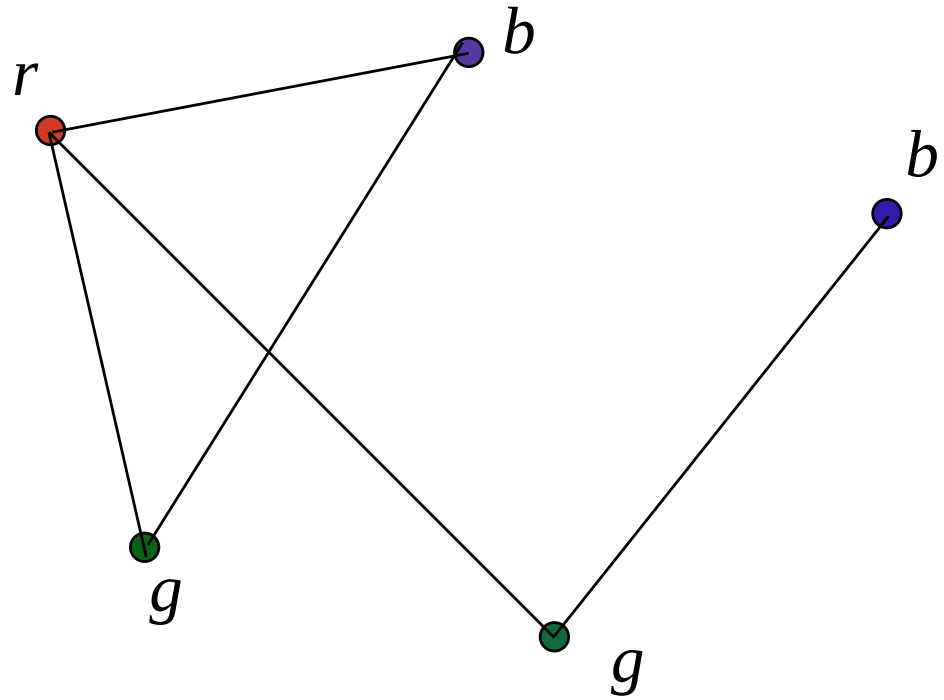
A regular k-coloring of the vertices of a graph G is a coloring of the vertices with k colors such that adjacent vertices have different colors.

The **chromatic number** $\chi(G)$ *of a graph G is the smallest number k such that there exists a regular k-coloring of the vertices of G*

Notation: We use the notion *almost always* as an abbreviation for the property that the probability that a certain condition holds converges to 1 as the *size* of the problem goes to the infinity.

Chromatic Number of Random Graphs

Chromatic Number $= 3$



Chromatic Number of Random Graphs

Theorem (Bollobás, Frieze, Grimmet, McDiarmid)

(i) If $C_0/n \le p = p(n) \le (\log n)^{-7}$ (for a proper constant $C_0 > 0$) then almost always

$$\frac{np}{2\log(np) - 2\log\log(np) + 1} \le \chi(G(n,p)) \le \frac{np}{2\log(np) - 40\log\log(np)}.$$

(ii) If $(\log n)^{-2} \le p = p(n) \le c$ (for some arbitrary $c < 1$) then almost always

$$\frac{n}{2\log_b n - \log_b\log_b n} \le \chi(G(n,p)) \le \frac{n}{\log_b n - 6\log_b\log_b n},$$

where $b = 1/(1-p)$.

(iii) If $p = p(n) > n^{-\delta}$ for every $\delta > 0$ (and sufficiently large n) but $p = p(n) \le c$ (for some arbitrary $c < 1$) then almost always

$$\chi(G(n,p)) = \frac{n}{2\log_b n - 2\log_b\log_b n + O(1/p)}.$$

Chromatic Number of Random Graphs

Theorem (Łuczak, Alon and Krivelevich 1997)

Fix some $\varepsilon > 0$. For every sequence $p = p(n)$ there exists a function $h(n)$ such that almost always

(i) if $p \geq n^{-\frac{1}{2}-\varepsilon}$ then $\chi(G(n,p)) \sim h(n)$, and

(ii) if $p \leq n^{-\frac{1}{2}-\varepsilon}$ then $\boxed{h(n) \leq \chi(G(n,p)) \leq h(n)+1}$.

Chromatic Number of Random Graphs

Theorem (Shamir and Spencer 1987)

$$\boxed{\mathbb{P}\{|\chi(G(n,p)) - \mathbb{E}\,(\chi(G(n,p)))| > \lambda\sqrt{n-1}\} < 2\,e^{-\lambda^2/2}.}$$

Remark. This theorem is weaker than the previous one (for $p \leq n^{-\frac{1}{2}-\epsilon}$) but the basis for further considerations.

Chromatic Number of Random Graphs

Definition. *A* **martingale** *is a sequence of random variables* $Y_0, Y_1 \ldots, Y_n$ *on a probabilty space* $(\Omega, \mathcal{F}, \mathbb{P})$ *with*

$$\boxed{\mathbb{E}\,(Y_{k+1}|\mathcal{F}_k) = Y_k,}$$

where $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n = \mathcal{F}$ *is an increasing sequence of* σ*-fields.*

Theorem (Azuma's Inequality) Suppose that $Y_0, Y_1 \ldots, Y_n$ is a martingale with constant Y_0 and that

$$\boxed{|Y_{k+1} - Y_k| \leq c_k}$$

for some some constants c_k $(0 \leq k < n)$. Then, for every $t > 0$,

$$\boxed{\mathbb{P}\{|Y_n - \mathbb{E}\,Y_n| \geq t\} \leq 2\,\exp\left(-\frac{t^2}{2\sum_{k=1}^{n} c_k^2}\right).}$$

Chromatic Number of Random Graphs

Theorem (McDiarmid's Independent Bounded Difference inequality)

Let $X_1, \ldots, X_n$ be independent random variables, with X_k taking values in a set Ω_k. Suppose that a function $f : \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ satisfies the property that

$$|f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)| \leq c_k$$

if $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ differ only at the k-th coordinate, that is $x_j = y_j$ for $j \neq k$.

Then, the random variable $Y = f(X_1, \ldots, X_n)$ satisfies, for any $t \geq 0$,

$$\boxed{\mathbb{P}\{|Y - \mathbb{E}\, Y| \geq t\} \leq 2 \exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right).}$$

Chromatic Number of Random Graphs

Proof

$\mathcal{F}_k$ ... σ-field generated by $X_1, \ldots, X_k$

$Y_k = \mathbb{E}\,(f(X_1, \ldots, X_n)|\mathcal{F}_k)$, $k = 0, 1, \ldots, n$ is a martingale

$|f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)| \leq c_k$ implies that $|Y_{k+1} - Y_k| \leq c_k$.

Hence, Azuma's inequality applies.

Chromatic Number of Random Graphs

Vertex Exposure Martingale

$A_k = \{\{j,k\} : 1 \leq j < k\}$ ... edges that connect k with $j < k$.

$X_k = (\mathbb{I}_{[e \in G(n,p)]} : e \in A_k)$ ... rand. vector of indicators of edges in A_k.

f ... *graph theoretical function* (for example, the chromatic number).

$\mathcal{F}_k$ ... σ-field generated by $X_1, \ldots, X_k$

$\boxed{Y_k = \mathbb{E}\,(f(G(n,p))|\mathcal{F}_k)}$ ... **vertex exposure martingale**

(It can be interpreted as the conditional expectation of f with partial information on the first k vertices and their internal edges.)

Chromatic Number of Random Graphs

Remark. $|f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)| \leq c_k$ with $x_k, y_k \in A_k$ says that $|f(H_1) - f(H_2)| \leq c_k$ if H_1, H_2 are subgraphs of the complete graph on the vertices $\{1, 2, \ldots, n\}$ such that the symmetric difference of the edge sets of H_1 and H_2 is contained in A_k.

If one adds a vertex to a graph then the chromatic number changes at most by 1. (Here we use the vertex k.)

$\Longrightarrow$ This condition is satisfied for the chromatic number with $c_k = 1$.

Travelling Salesman Problem

$\mathbf{X} = (X_1, X_2, \ldots, X_n)$ ... n-tuple of random point selected uniformly and independently in the unit square $[0,1]^2$

Length of the minimum (travelling salesman) tour:

$$\boxed{\mathrm{TSP}(\mathbf{X}) = \min_{\pi \in S_n} \sum_{j=1}^{n} \left| X_{\pi(j)} - X_{\pi(j+1)} \right|}$$

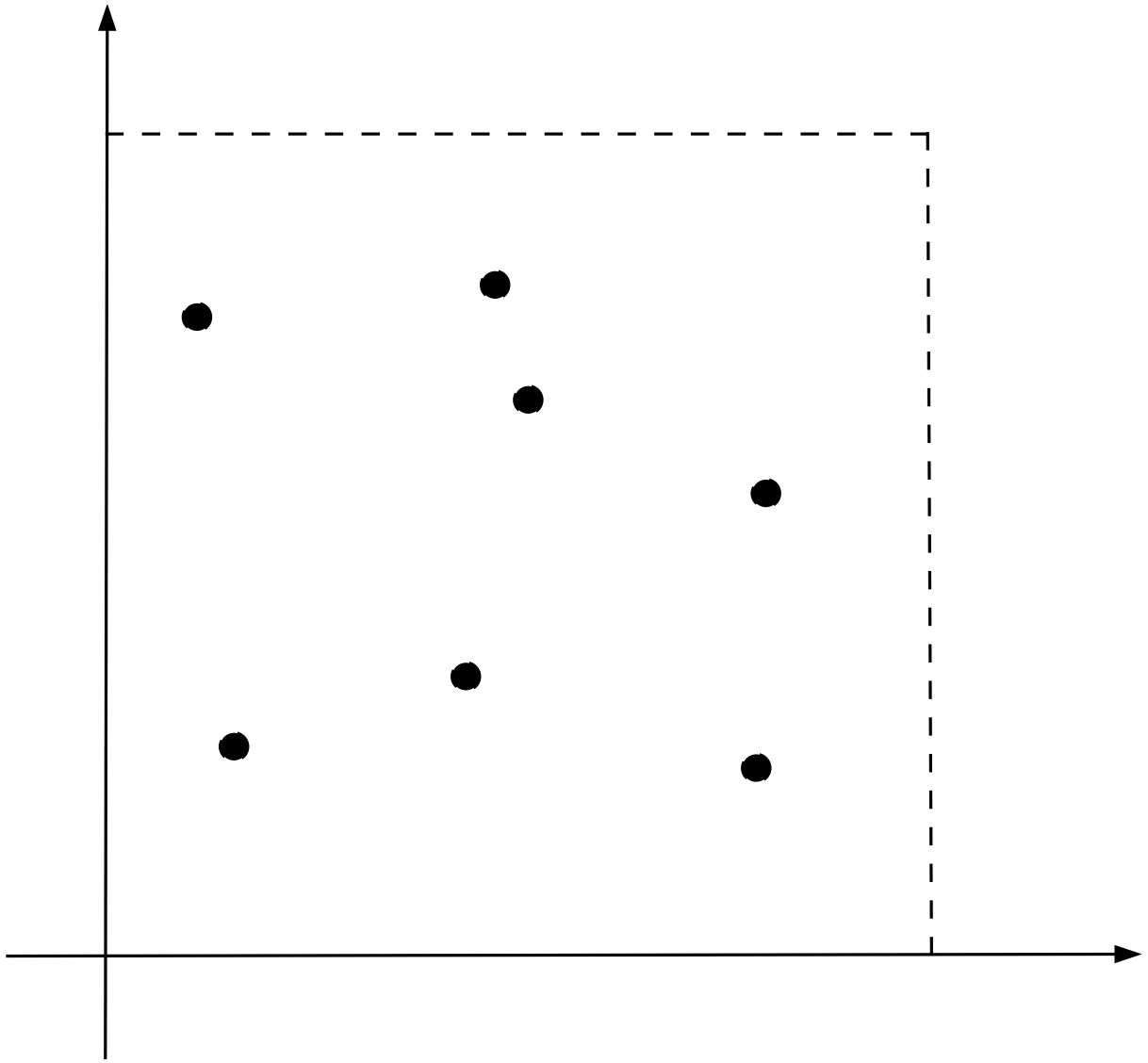
Theorem (Beardwood, Halton and Hammersley 1959)

$$\boxed{\frac{\mathrm{TSP}(\mathbf{X})}{\sqrt{n}} \to \beta_2} \quad \text{in prob.}$$

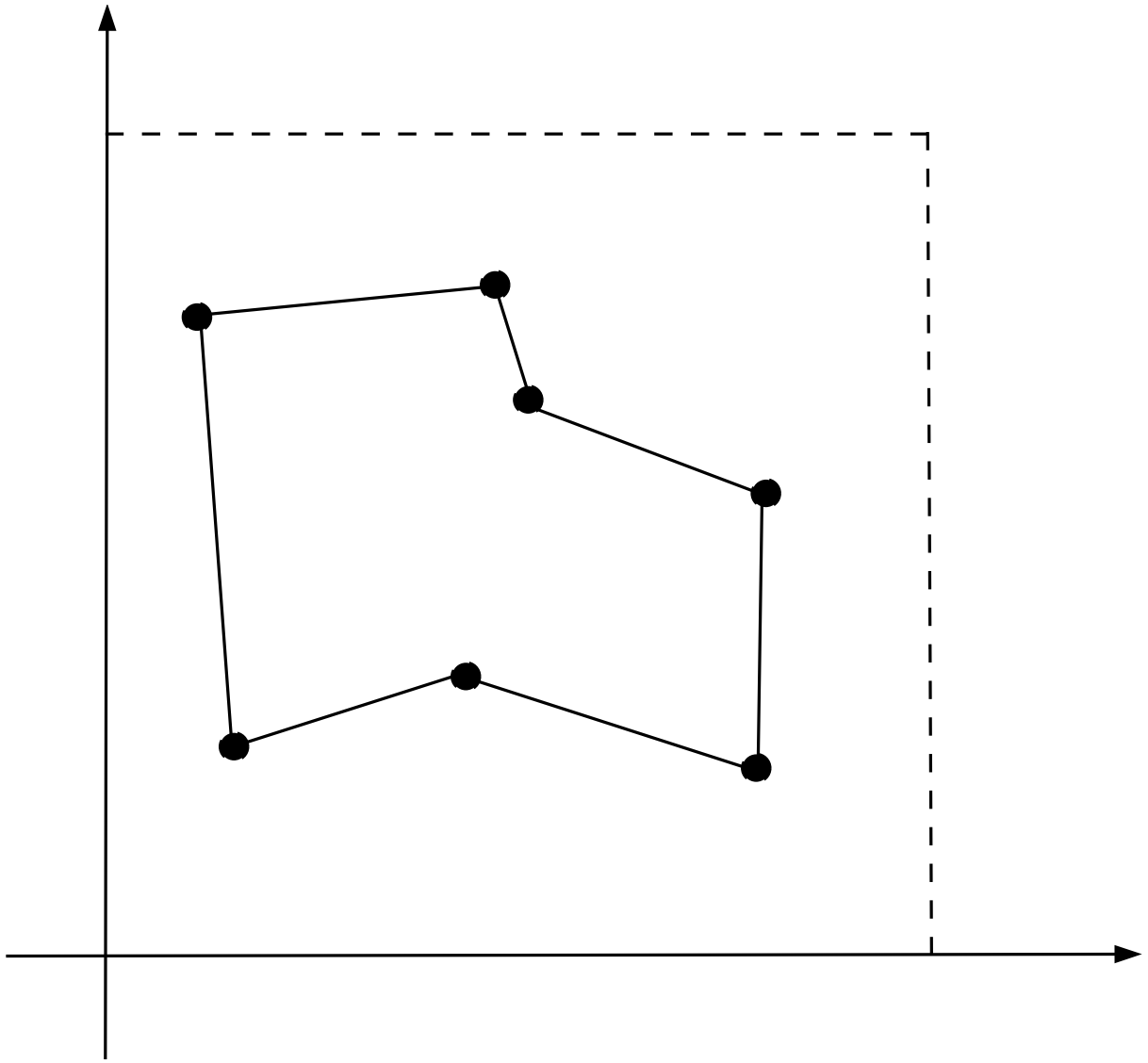
for some $\beta_2 > 0$.

Remark: Up to now there is no known analytic expression for β_2.

Travelling Salesman Problem



Travelling Salesman Problem



Travelling Salesman Problem

Notation: M(Y) ... median of r.v. Y

Theorem (Rhee and Talagrand)

$$\boxed{\mathbb{P}\left\{|\mathrm{TSP}(\mathbf{X}) - \mathrm{M}(\mathrm{TSP}(\mathbf{X}))| \geq t\right\} < 4e^{-t^2/c}.}$$

for some constant $c > 0$.

Corollary. All central moments of TSP($\mathbf{X}$) are bounded.

(However, the exact location of the mean is unknown.)

Travelling Salesman Problem

Talagrand's Inequality

$\Omega_1, \Omega_2, \ldots, \Omega_n \ldots$ probability spaces, $\Omega = \Omega_1 \times \cdots \times \Omega_n$

$\mathbf{X} = (X_1, X_2, \ldots, X_n)$ ... independent random variables, X_k taking values in Ω_k.

Weighted Hamming distance related to $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $\alpha_k \geq 0$:

$$d_\alpha(\mathbf{x}, \mathbf{y}) = \sum_{x_i \neq y_i} \alpha_i$$

Talagrand's convex distance

$$\boxed{d_T(\mathbf{x}, A) = \sup_{\alpha \geq 0, \|\alpha\|=1} \inf_{\mathbf{y} \in A} d_\alpha(\mathbf{x}, \mathbf{y})}$$

between $\mathbf{x} \in \Omega$ and $A \subset \Omega$.

Travelling Salesman Problem

Talagrand's Inequality

$$\boxed{\mathbb{P}\{\mathbf{X} \in A\} \cdot \mathbb{P}\{d_T(\mathbf{X}, A) \geq t\} < e^{-t^2/4}.}$$

Travelling Salesman Problem

Theorem

f ... real valued function on $\Omega = \Omega_1 \times \cdots \times \Omega_n$

For every $\mathbf{x} \in \Omega$ there exists a non-negative unit n-vector α and a constant $c > 0$ such that for all $\mathbf{y} \in \Omega$

$$\boxed{f(\mathbf{x}) \leq f(\mathbf{y}) + c\, d_\alpha(\mathbf{x}, \mathbf{y}).}$$

Then, for every random n-tuple $\mathbf{X} = (X_1, \ldots, X_n)$ of independent random variables X_k taking values in Ω_k we have

$$\boxed{\mathbb{P}\{|f(\mathbf{X}) - \mathsf{M}(f(\mathbf{X}))| \geq t\} \leq 4\, e^{-t^2/(4c^2)}.}$$

Travelling Salesman Problem

Proof

$A_a := \{\mathbf{y} \in \Omega : f(\mathbf{y}) \leq a\}$

By assumption for every $\mathbf{x} \in \Omega$ there exists a non-negative unit n-vector α such that for all $\mathbf{y} \in A_a$:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + c\, d_\alpha(\mathbf{x}, \mathbf{y}) \leq a + c\, d_\alpha(\mathbf{x}, \mathbf{y}).$$

By taking the miminum over all $\mathbf{y} \in A_a$ we, thus, get

$$f(\mathbf{x}) \leq a + c\, d_\alpha(\mathbf{x}, A_a) \leq a + c\, d_T(\mathbf{x}, A_a).$$

Hence

$$\boxed{f(\mathbf{x}) \geq a + t \quad \Longrightarrow \quad d_T(\mathbf{x}, A_a) \geq t/c.}$$

$$\begin{aligned} \Longrightarrow \quad \mathbb{P}\{f(\mathbf{X}) \leq a\} \cdot \mathbb{P}\{f(\mathbf{X}) \geq a + t\} &\leq \mathbb{P}\{\mathbf{X} \in A_a\} \cdot \mathbb{P}\{d_T(\mathbf{x}, A_a) \geq t/c\} \\ &\leq e^{-t^2/(4c^2)}. \end{aligned}$$

Travelling Salesman Problem

$a = \mathsf{M}(f(\mathbf{X}))$, $\mathbb{P}\{f(\mathbf{X}) \leq a\} = \frac{1}{2}$:

$$\mathbb{P}\{f(\mathbf{X}) \geq \mathsf{M}(f(\mathbf{X})) + t\} \leq 2\, e^{-t^2/(4c^2)}.$$

$a = \mathsf{M}(f(\mathbf{X})) - t$:

$$\mathbb{P}\{f(\mathbf{X}) \leq \mathsf{M}(f(\mathbf{X})) - t\} \leq 2\, e^{-t^2/(4c^2)}.$$

Travelling Salesman Problem

Lemma

For every $\mathbf{x} \in ([0,1]^2)^n$ there exists non-negative unit vector α and a constant $c > 0$ such that for all $\mathbf{y} \in ([0,1]^2)^n$

$$\boxed{\mathrm{TSP}(\mathbf{x}) \leq \mathrm{TSP}(\mathbf{y}) + c\, d_\alpha(\mathbf{x}, \mathbf{y}).}$$

(Elementary proof that uses an approximate minumum tour to construct α.)

Remark

This method can be applied to several other problem, for example to the minimal Steiner tree problem etc.

Longest Increasing Subsequence in R. P.

S_n ... the set of permutations of the numbers $\{1, 2, \ldots, n\}$
(We assume that every permutation in S_n is equally likely.)

For $\pi \in S_n$ we say that $\pi(i_1), \pi(i_2), \ldots, \pi(i_k)$ is an increasing subsequence in π if $i_1 < i_2 < \cdots < i_k$ and $\pi(i_1) < \pi(i_2) < \cdots < \pi(i_k)$.

$\boxed{L_n = L_n(\pi) \ \ldots \text{ length of the longest increasing subsequence.}}$

Ulam's Problem: $\mathbb{E}\, L_n \sim$?

Ulam's conjecture: $\boxed{\mathbb{E}\, L_n \sim c\sqrt{N}}$ for some constant $c > 0$.

Longest Increasing Subsequence in R. P.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 6 & 2 & 4 \end{pmatrix}$$

Longest Increasing Subsequence in R. P.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 5 & 6 & 2 & 4 \end{pmatrix}$$

Longest Increasing Subsequence in R. P.

Erdős Szekeres 1935: $c \geq \frac{1}{2}$

Logan and Shepp 1977: $c \geq 2$

Vershik and Kerov 1977: $\boxed{c = 2}$.

(Alternate proofs are due to Aldous and Diaconis, Seppäläinen, and Johansson).

Frieze 1991: L_n is concentrated

Bollobás and Brightwell 1992, Talagrand 1995: $\mathbb{V}L_n = O(\sqrt{N})$.

Odlyzko and Rains 2000: order of $\mathbb{V}L_n$ should be $N^{1/3}$.

Baik, Deift, and Johansson 1999: complete solution.

Longest Increasing Subsequence in R. P.

Theorem (Baik, Deift, and Johansson 1999)

Let S_n be the group of permutations of n numbers with uniform distribution and L_n the longest increasing subsequence. Then there exists a random variable Y such that

$$\boxed{\frac{L_n - 2\sqrt{n}}{n^{1/6}} \xrightarrow{\mathsf{d}} Y.}$$

Furthermore, we have convergence of all moments.

Remark.

The limiting distribution Y is exactly the same at the limiting distribution of the largest eigenvalue in random Hermitian matrices. However, it seems that there is no direct connection between these two problems.

Longest Increasing Subsequence in R. P.

Tracy-Widom distribution: $F(t) = \mathbb{P}\{Y \leq t\}$

$u(x)$ ... solution of the Painlevé II equation

$$u'' = 2u^3 + xu, \quad u(x) \sim Ai(x) \quad (\text{as } x \to \infty);$$

$Ai(x)$ denotes the Airy function.

$$F(t) = \exp\left(\int_t^\infty (x-t)^2 u(x)^2 \, dx\right),$$

Longest Increasing Subsequence in R. P.

Proof Method

Basically one determines the asymptotic behaviour of the Poisson transform

$$\phi_k(\lambda) = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!}\mathbb{P}\{L_n \leq k\}$$

that can be represented as

$$\phi_k(\lambda) = \frac{e^{-\lambda}}{(2\pi)^k k!}\int_{[-\pi,\pi]^k} \exp\left(2\sqrt{\lambda}\sum_{j=1}^{k}\theta_j\right) \prod_{1\leq j<\ell\leq k}\left|e^{i\theta_j} - e^{i\theta_\ell}\right| d\theta_1 \cdots d\theta_k.$$

One has to use the theory of orthogonal polynomials on the unit circle, sophisticated Riemann-Hilbert problem techniques and certain properties on eigenvalues of random matrices.

Diameter and Maximum Degree in R. Graphs

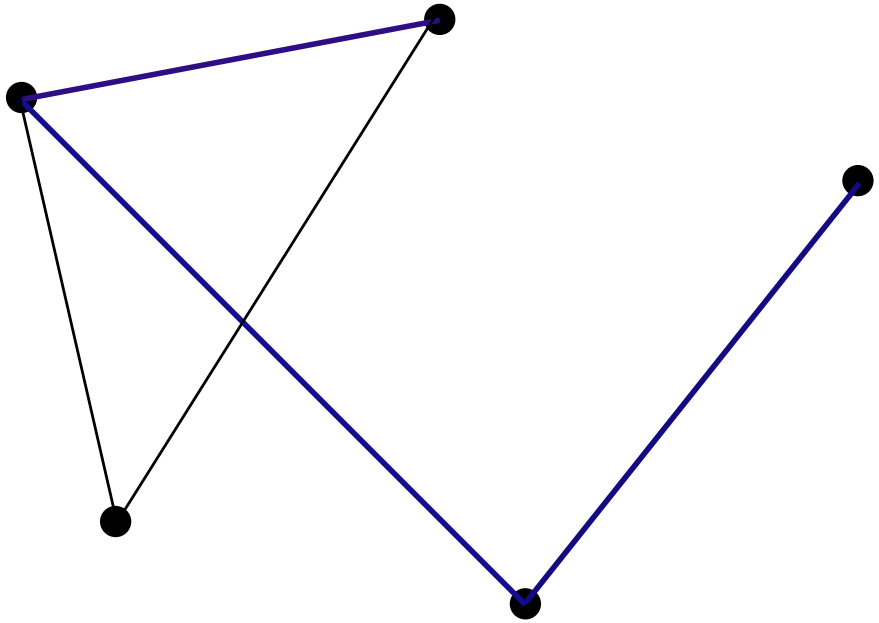
Definition

The **diameter** $\boxed{\text{diam}(G)}$ *of a graph connected* G *is the largest distance between two nodes in* G*. If* G *is not connected then* $\text{diam}(G) = \infty$*.*

The **maximum degree** *of an (undirected) graph* G *will be denoted by* $\boxed{\Delta(G)}$*.*

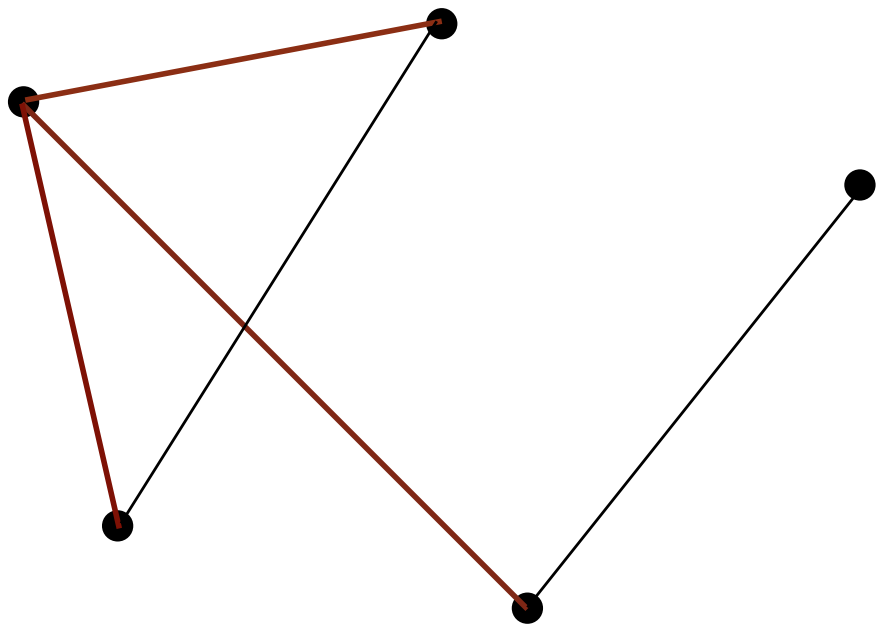
Diameter and Maximum Degree in R. Graphs

Diameter $= 3$



Diameter and Maximum Degree in R. Graphs

Maximum Degree $= 3$



Diameter and Maximum Degree in R. Graphs

$G(n,p)$-**Random Graphs**

Theorem (Burtin, Bollobás)

(i) If $(pn)/\log n \to \infty$ and $\log n/\log(pn) \to \infty$ then almost always

$$\text{diam}(G(n,p)) \sim \frac{\log n}{\log(pn)}.$$

(ii) Let c be a positive constant and $p = p(n)$ and $d = d(n)$, an integer ≥ 2, be related by $p^d n^{d-1} = \log(n^2/c)$. Further suppose that $(pn)/(\log n)^3 \to \infty$. Then

$$\lim_{n\to\infty} \mathbb{P}\{\text{diam}(G(n,p)) = d\} = e^{-c/2}$$

and

$$\lim_{n\to\infty} \mathbb{P}\{\text{diam}(G(n,p)) = d+1\} = 1 - e^{-c/2}.$$

Diameter and Maximum Degree in R. Graphs

Notation

$$\lambda_k = \lambda_k(n) := n\binom{n-1}{k} p(n)^k (1-p(n))^{n-k-1}.$$

$k = k(n) \geq np$ be such that the quantity $\max\{\lambda_k, 1/\lambda_k\}$ is minimal.

Diameter and Maximum Degree in R. Graphs

Theorem (Bollobas) Suppose that $p = p(n) = o(\log n/n)$.

(i) If $0 < \liminf \lambda_k \leq \limsup \lambda_k < \infty$ then, as $n \to \infty$,

$$\mathbb{P}\{\Delta = k(n)\} = 1 - e^{-\lambda_k} + o(1)$$

and

$$\mathbb{P}\{\Delta = k(n)\} = e^{-\lambda_k} + o(1).$$

(ii) If $\lim \lambda_k = \infty$ then

$$\mathbb{P}\{\Delta = k(n)\} = 1 + o(1).$$

(iii) If $\lim \lambda_k = 0$ then

$$\mathbb{P}\{\Delta = k(n) - 1\} = 1 + o(1).$$

(iv) If there is a function $D(n)$ with $\mathbb{P}\{\Delta = D(n)\} = 1 + o(1)$ as $n \to \infty$ then $p = p(n) = o(\log n/n)$.

Diameter and Maximum Degree in R. Graphs

Theorem (Bollobas, Riordan and Selby)

Suppose that $0 < p < 1$ is fixed and $q = 1 - p$.

(i) For every real number y we have

$$\mathbb{P}\left\{\Delta \leq pn + \sqrt{2pqn \log n}\left(1 - \frac{\log\log n}{4\log n} + \frac{y - 2\sqrt{\pi}}{2\log n}\right)\right\} = \exp\left(-e^{-y}\right) + o(1).$$

(ii) Almost always we have

$$\left|\Delta - pn - \sqrt{2pqn\log n} + \log\log n\sqrt{\frac{pqn}{8\log n}}\right| \leq \log\log n\sqrt{\frac{n}{\log n}}.$$

(iii) For every real number b there exists $c(b)$ such that

$$\mathbb{P}\{\Delta < pn + b\sqrt{npq}\} = (c(b) + o(1))^n.$$

Diameter and Maximum Degree in R. Graphs

Barabási-Albert model (for real-world graphs, internet etc.):

- Randomly growing graph
- A new node is joint to an existing one with probability proportional to the degree.

This definition is not unambigous!!!!

Diameter and Maximum Degree in R. Graphs

Scale-Free Random Graphs

A *power law* is a distribution Z with tail of the form $\mathbb{P}\{Z = d\} \sim c \cdot d^{-k}$ (for some $k > 1$).

If a (random graph) that has an power law as (asymptotic) degree distribution is called **scale-free**.

Diameter and Maximum Degree in R. Graphs

Bollobás and Riordan: G_m^n multi-graph

$m = 1$ (G_1^n):

- Initial node 1 with a loop.

- at step k we add one node that is connected to $j \leq k$ with propability

$$\begin{cases} \dfrac{\deg_{G_1^{k-1}}(j)}{2k-1} & \text{if } j < k, \\ \dfrac{1}{2k-1} & \text{if } j = k. \end{cases}$$

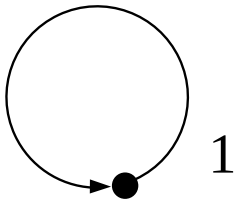
Diameter and Maximum Degree in R. Graphs

$m \geq 1$

- G_m^n is constructed from G_1^{mn} by identifying the nodes $\{(\ell-1)m+1, (\ell-1)m+2, \ldots, \ell m\}$ $(1 \leq \ell \leq n)$ of G_1^{mn} to a new node ℓ (and all edged within the nodes $\{(\ell-1)m+1, (\ell-1)m+2, \ldots, \ell m\}$ are now loops of the new node ℓ)

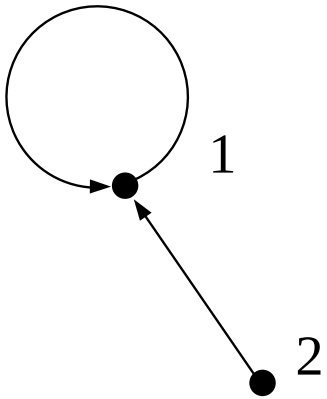
Diameter and Maximum Degree in R. Graphs

G_1^n



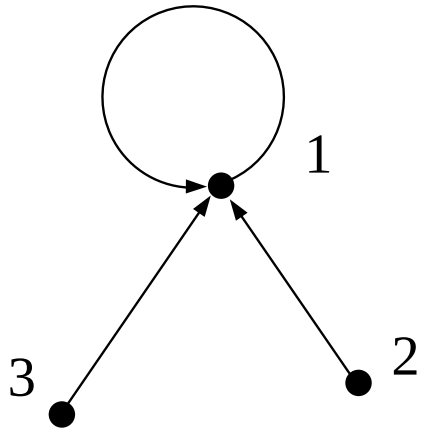
Diameter and Maximum Degree in R. Graphs

G_1^n



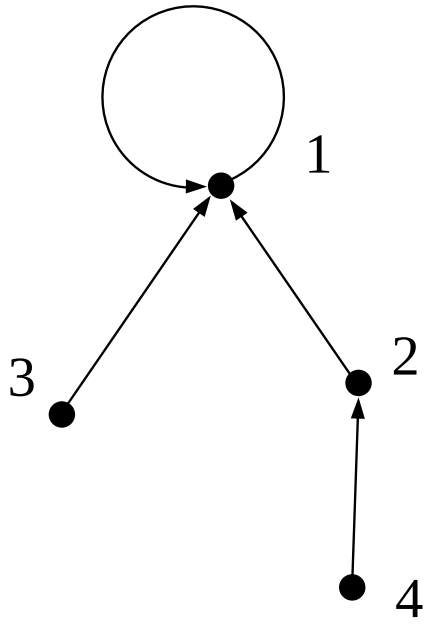
Diameter and Maximum Degree in R. Graphs

G_1^n



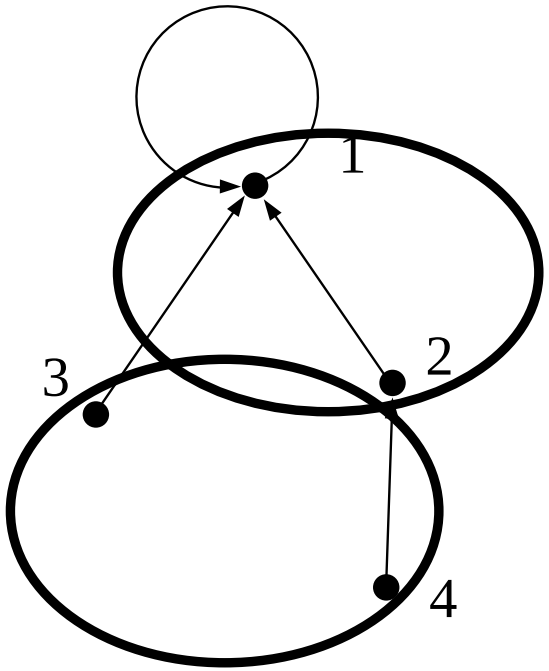
Diameter and Maximum Degree in R. Graphs

G_1^n



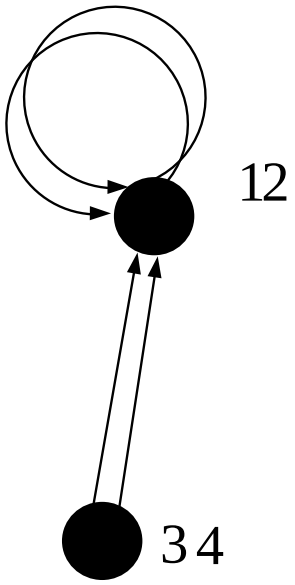
Diameter and Maximum Degree in R. Graphs

G_2^n



Diameter and Maximum Degree in R. Graphs

G_2^n



Diameter and Maximum Degree in R. Graphs

Scale-Free Random Graphs: G_m^n is scale free, the tail of the degree distribution is of the form $c \cdot d^{-3}$.

Theorem (Bollobás and Riordan 2004)

Suppose that $m \geq 2$. Then for every $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}\left\{(1-\epsilon)\frac{\log n}{\log \log n} \leq \text{diam}(G_m^n) \leq (1+\epsilon)\frac{\log n}{\log \log n}\right\} = 1.$$

Height and Maximum Degree of Random Trees

- Galton-Watson Trees
- Pólya Trees
- m-Ary Search Trees
- Recursive Trees
- Scale Free Trees
- Tries
- Digital Search Trees

Galton-Watson Trees

ξ ... non-negative integer valued random variable,

$$\mathbb{E}\,\xi = 1,\ 0 < \mathbb{V}\,\xi = \sigma^2 < \infty.$$

$(Z_k)_{k\geq 0}$ Galton-Watson branching process: $Z_0 = 1$,

$$\boxed{Z_k = \sum_{j=1}^{Z_{k-1}} \xi_j^{(k)},}$$

where the $(\xi_j^{(k)})_{k,j}$ are iid random variables distributed as ξ.

A Galton-Watson branching processes can be represented by ordered (finite or infinite) rooted trees T.

$y_n = \mathbb{P}\{|T| = n\}$, $y(x) = \sum_{n\geq 1} y_n x^n$, $\varphi(t) = \mathbb{E}\,t^{\xi}$:

$$\Longrightarrow \quad \boxed{y(x) = x\varphi(y(x)).}$$

Galton-Watson Trees

$\mathcal{T}_n$ ... set of rooted trees T of size $|T| = n$,
$\nu(T)$ ... probabilty that T occurs in Galton-Watson branching process:

$$\nu_n(T) := \frac{\nu(T)}{y_n}$$

is a probability distribution on $\mathcal{T}_n$.

$\sigma^2 < \infty$:

$$y_n \sim \frac{d}{\sqrt{2\pi}\sigma} n^{-3/2} \qquad (n \equiv 1 \bmod d),$$

where $d = \gcd\{i > 0 : \mathbb{P}\{\xi = i\} > 0\}$.

Galton-Watson Trees

Examples.

$\varphi(t) = \mathbb{E}\, t^{\xi} = (1+t)^2/4 = \frac{1}{4} + \frac{t}{2} + \frac{t^2}{4}$: **binary trees** with n (internal) nodes, where each binary tree (of size n) has equal probability.

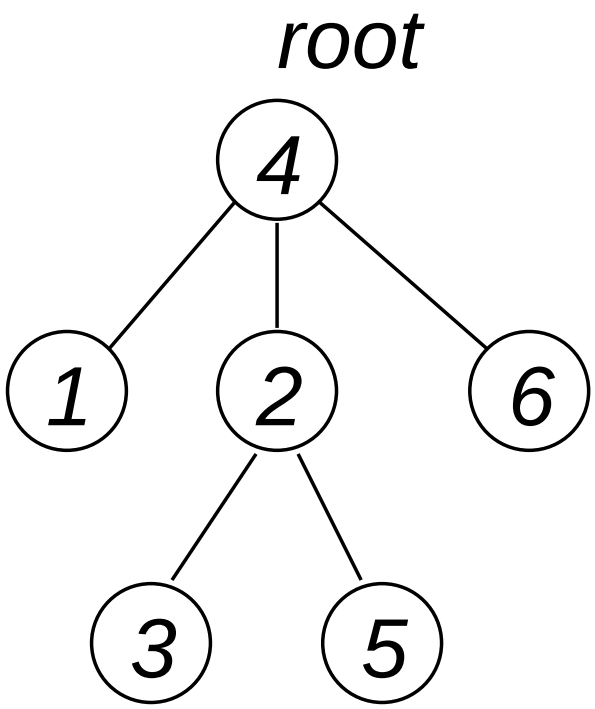
$\varphi(t) = \mathbb{E}\, t^{\xi} = 1/(2-t) = \frac{1}{2} + \frac{t}{4} + \frac{t^2}{8} + \cdots$: **planted plane trees**, every rooted planar tree is equally likely.

$\varphi(t) = e^{t-1}$: **Cayley trees**, every rooted labeled tree is equally likely.

Equivalent description: Simply generated trees (introduced by Meir and Moon)

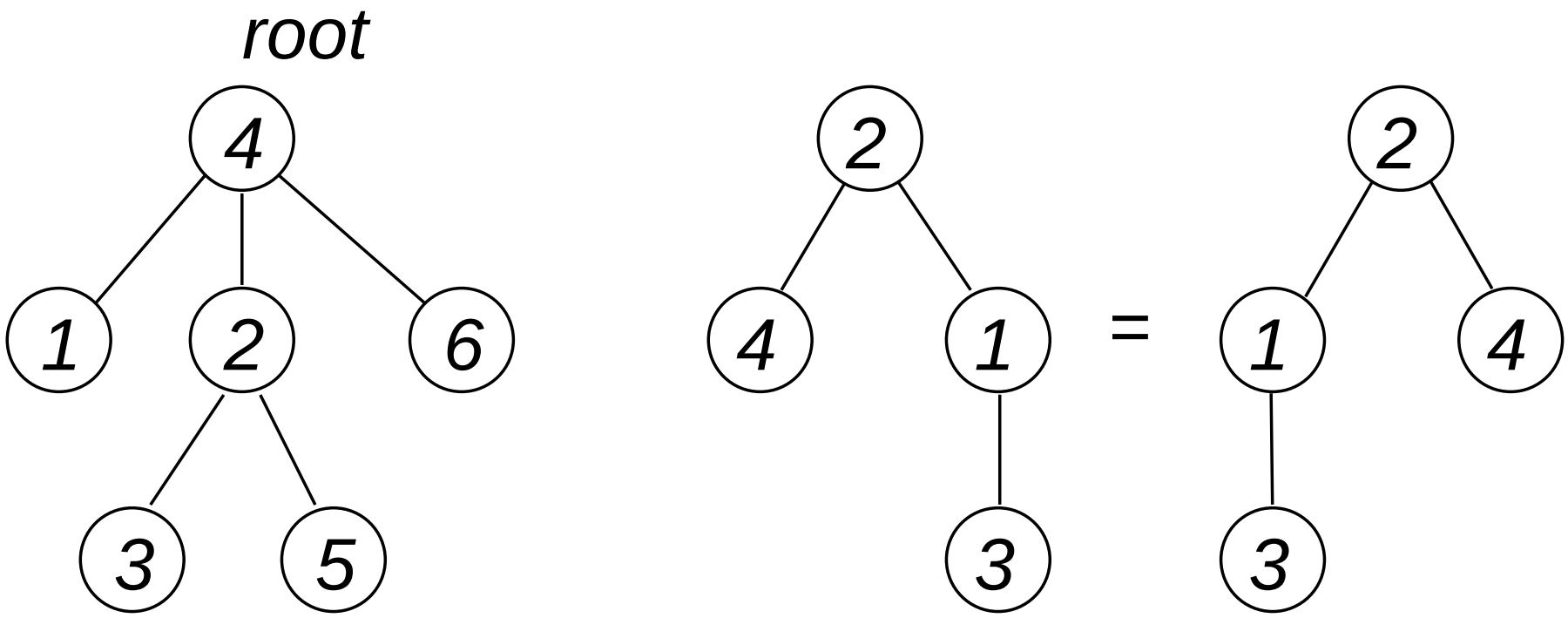
Galton-Watson Trees

Cayley Trees: labeled, rooted, non-planar



Galton-Watson Trees

Cayley Trees: labeled, rooted, **non-planar**



Galton-Watson Trees

Theorem (de Bruijn, Knuth and Rice, Flajolet, Gao, Odlyzko and Richmond, Aldous)

Suppose that the second moment $\mathbb{E}\,\xi^2$ is finite. Then

$$\boxed{\frac{1}{\sqrt{n}} H_n \xrightarrow{\mathrm{d}} \frac{2}{\sigma} \max_{0 \le t \le 1} e(t),}$$

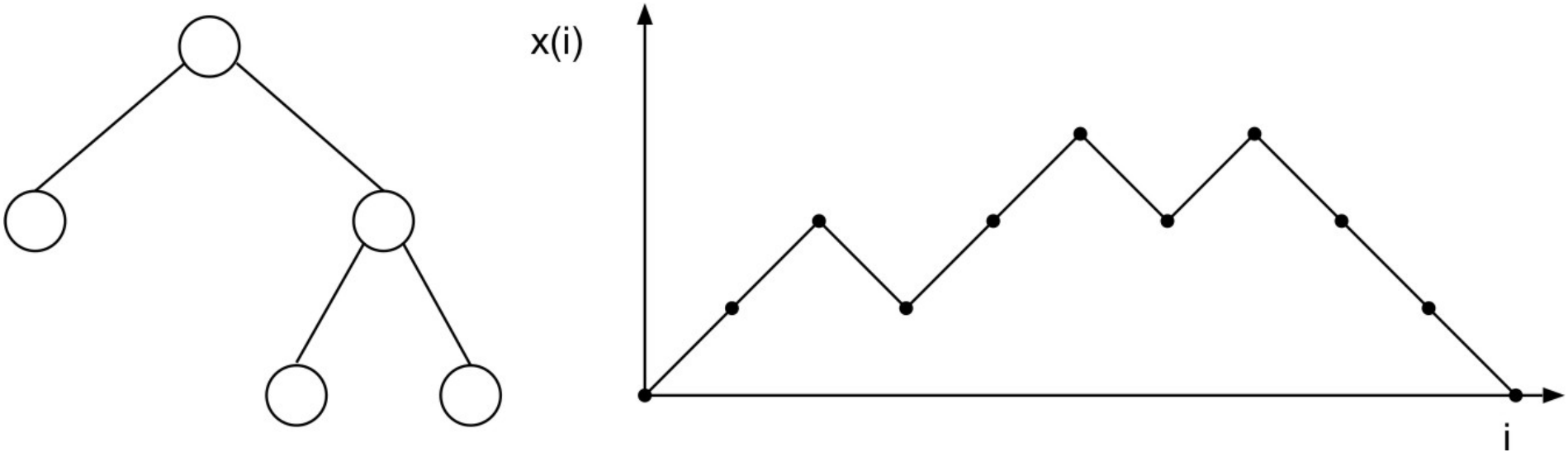
where $(e(t), 0 \le t \le 1)$ denotes Brownian excursion of duration 1. Furthermore, if $\varphi(t) = \mathbb{E}\, t^{\xi}$ exists for some $t > 1$ then we also have convergence of all moments. For every $r \ge 0$ we have, as $n \to \infty$,

$$\boxed{\mathbb{E}\,(H_n^r) = 2^{r/2} \sigma^{-r} r(r-1)\Gamma(r/2)\zeta(r) \cdot n^{r/2} \left(1 + O(n^{-\frac{1}{4}+\eta})\right)}$$

where $\zeta(s)$ denotes the Riemann Zeta-function and $(r-1)\zeta(r) = 1$ for $r = 1$ and η is any positive number.

Galton-Watson Trees

Depth-first search.



$$\frac{T_n(\lfloor 2nt \rfloor)}{c_2\sqrt{n}} \to e(t) \quad \ldots \text{ Brownian excursion}$$

Galton-Watson Trees

h_T ... height of T

$$y_k(x) = \sum_{n\geq 1} \mathbb{P}\{|T| = n,\, h_T \leq k\}\, x^n,$$

$$\Longrightarrow \quad \boxed{y_{k+1}(x) = x\varphi(y_k(x))}, \qquad y_0(x) = \varphi_0 x.$$

$$H(x) := \sum_{n\geq 1} \mathbb{E}\ H_n \cdot y_n \cdot x^n = \sum_{k\geq 0} (y(x) - y_k(x)).$$

A subtle analysis of the above recurrence yiels

$$H(x) = \frac{1}{\sigma^2} \log \frac{1}{1-x} + K + O\left(|1-x|^{\frac{1}{4}-\eta}\right)$$

for some constant K and every (fixed) $\eta > 0$

$$\Longrightarrow \quad \boxed{\mathbb{E}\ H_n = \frac{\sqrt{2\pi}}{\sigma} \cdot \sqrt{n} + O\left(n^{\frac{1}{4}+\eta}\right).}$$

Galton-Watson Trees

Theorem (Meir and Moon, Carr, Goh and Schmutz)

(i) Suppose that $\varphi_i = \Pr\{\xi = i\} > 0$ for sufficiently large $i \geq i_0$ and that $\varphi_{i+1}/\varphi_i \to 0$ as $i \to \infty$. Then

$$\mathbb{P}\{|\Delta(T_n) - \delta(n)| \leq 1\} = 1 + o(1),$$

where $\delta(n) = \max\{k \geq 0 : \mathbb{P}\{\xi \geq k\} \geq 1/n\}$.

(ii) If $\varphi(t) = e^{t-1}$ then there exists a sequence $\delta'(n)$ that is asymptotically equivalent to $\delta'(n) \sim \frac{\log n}{\log\log n}$ such that

$$\mathbb{P}\{\delta'(n) \leq \Delta(T_n) \leq \delta'(n) + 1\} = 1 + o(1).$$

(iii) If $\varphi(t) = 1/(2-t)$ then we have uniformly for all $k \geq 0$

$$\mathbb{P}\{\Delta(T_n) \leq k\} = \exp\left(-2^{-(k-\log_2 n+1)}\right) + o(1)$$

Galton-Watson Trees

$$y_d(x) = \sum_{n \geq 1} \mathbb{P}\{|T| = n,\, \Delta(T_n) \leq d\}\, x^n.$$

$$\Longrightarrow \quad \boxed{y_d(x) = x\, \varphi_d(y_d(x))} \quad \text{with} \quad \varphi_d(t) = \sum_{i \leq d} \varphi_i t^i.$$

$$\Longrightarrow \quad \mathbb{P}\{|T| = n,\, \Delta(T_n) \leq d\} \sim C_d (\varphi_d'(\tau_d))^n n^{-3/2},$$

where $\tau_d > 0$ is determined by $\tau_d \varphi_d'(\tau_d) = \varphi_d(\tau_d)$.

$$\Longrightarrow \quad \boxed{\mathbb{P}\{\Delta(T_n) \leq d\} = \frac{\mathbf{Pr}\{|T| = n,\, \Delta(T_n) \leq d\}}{\mathbf{Pr}\{|T| = n\}} \sim \sqrt{2\pi}\sigma C_d\, (\varphi_d'(\tau_d))^n.}$$

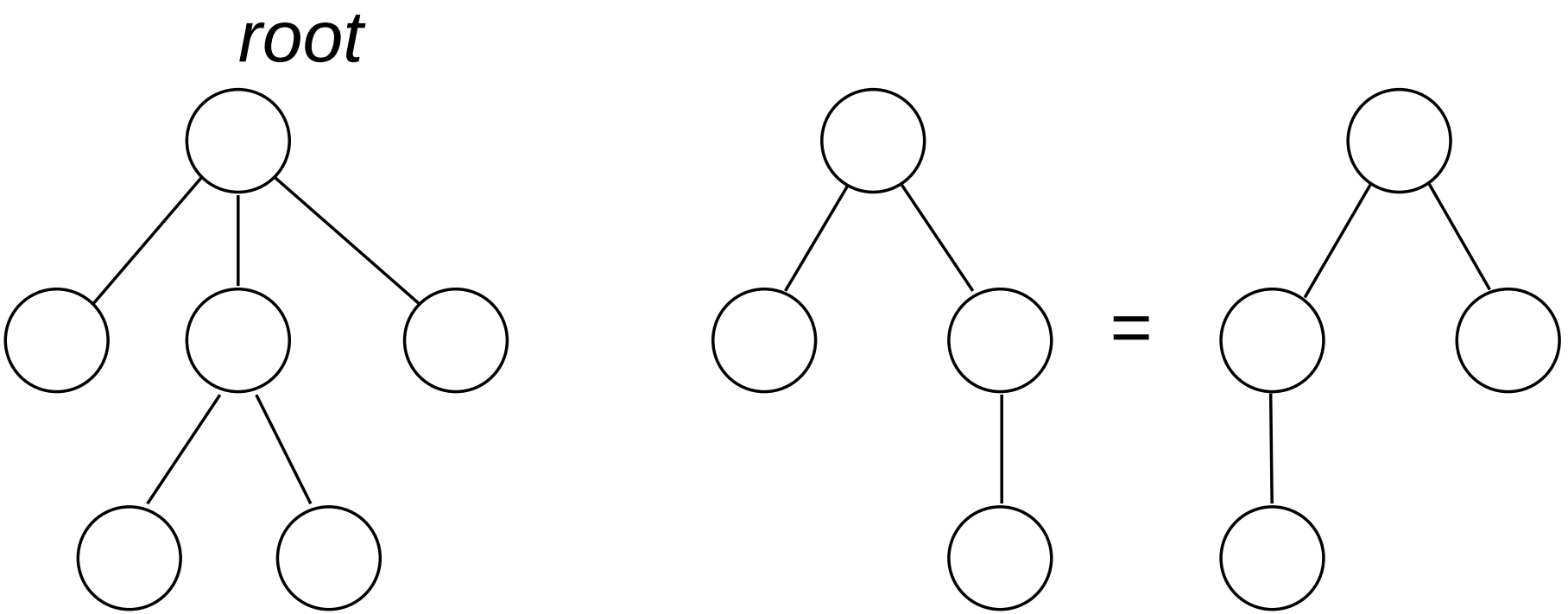
Pólya Trees

t_n ... number of rooted unlabeled (non-planar) trees

$$t(x) := \sum_{n \geq 1} t_n x^n.$$

$$\Longrightarrow \quad \boxed{t(x) = x \exp\left(t(x) + \frac{1}{2} t(x^2) + \frac{1}{3} t(x^3) + \cdots\right).}$$

Polya Trees



$$t(x) = \sum_{n \geq 1} t_n \, x^n \qquad t(x) = x \, e^{t(x) + \frac{1}{2} t(x^2) + \frac{1}{3} t(x^3) + \cdots}$$

Pólya Trees

The **height** of Pólya Trees has the same properties as Galton-Watson trees.

Theorem (Goh and Schmutz)

Let $\Delta(T_n)$ denote the maximum out-degree of Pólya trees of size n. Then

$$\boxed{\mathbb{P}\{\Delta(T_n) \leq k\} = \exp\left(-c_0 \eta^{k-\mu_n}\right) + o(1)}$$

with $c_0 = 3.262\ldots$, $\eta = 0.3383\ldots$, and $\mu_n = 0.9227\ldots \cdot \log n$.

Binary Search Trees

Storing Data

4,6,3,5,1,8,2,7

Binary Search Trees

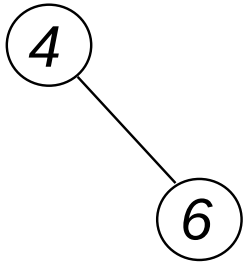
Storing Data

6,3,5,1,8,2,7

Binary Search Trees

Storing Data

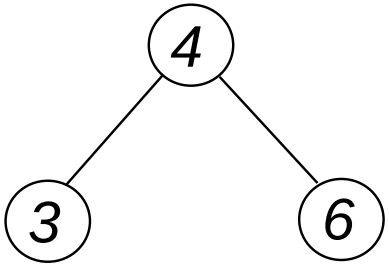
3,5,1,8,2,7



Binary Search Trees

Storing Data

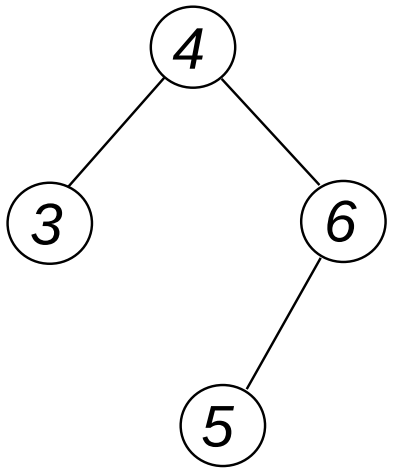
5,1,8,2,7



Binary Search Trees

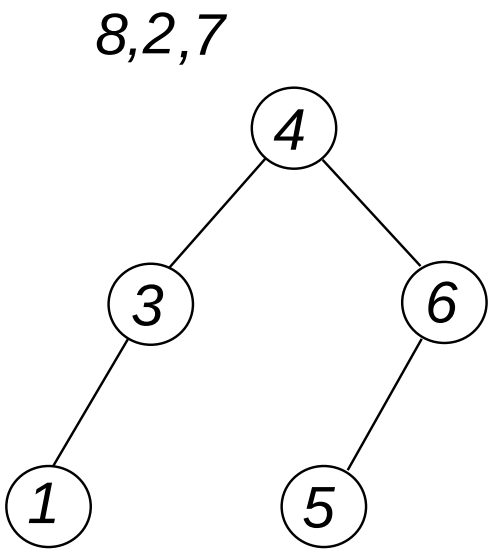
Storing Data

1,8,2,7



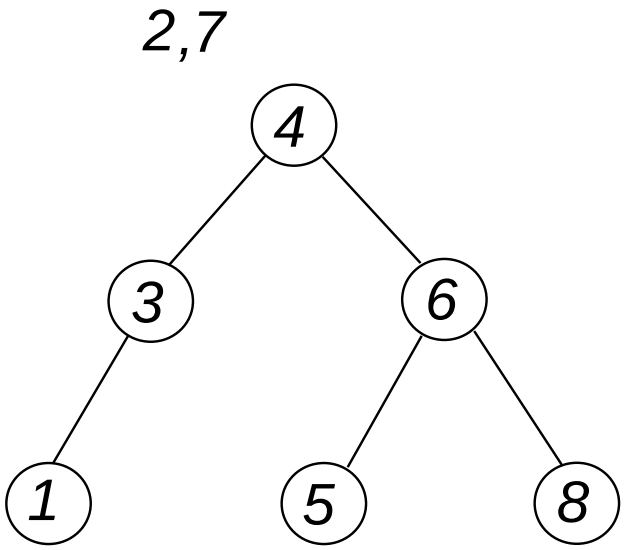
Binary Search Trees

Storing Data



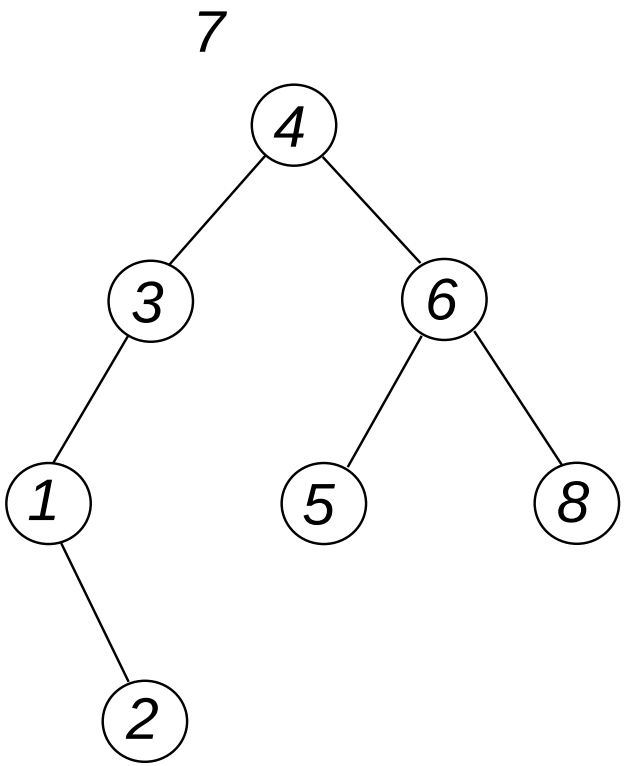
Binary Search Trees

Storing Data



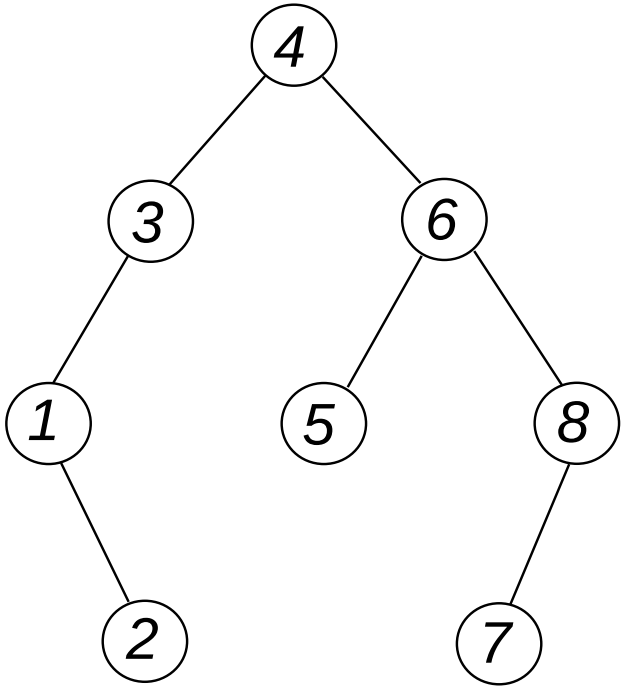
Binary Search Trees

Storing Data



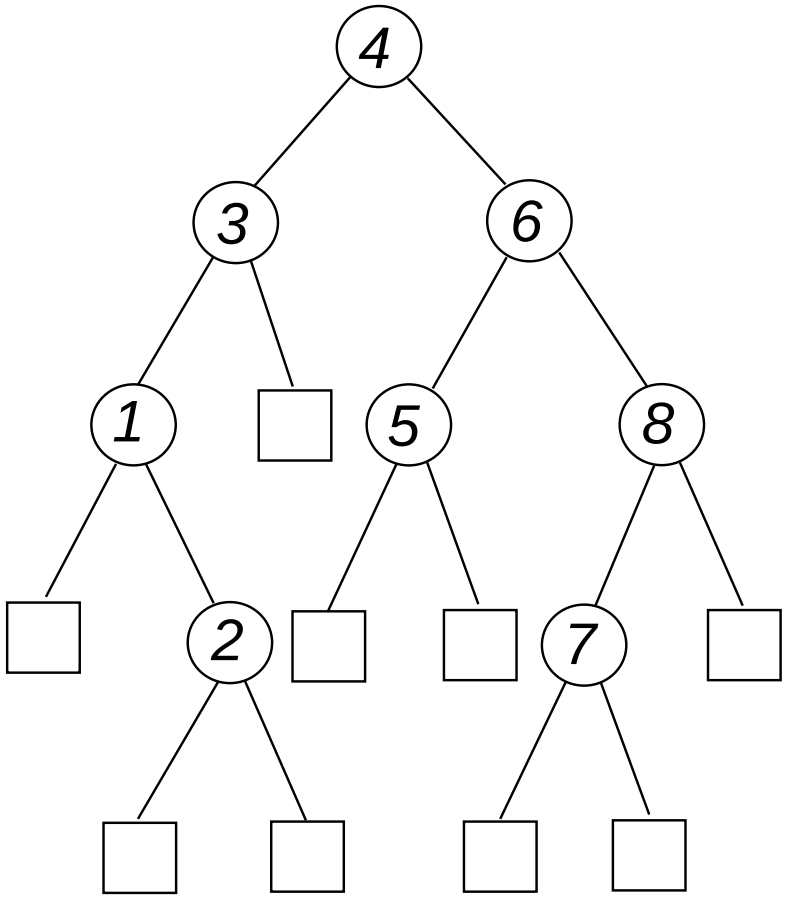
Binary Search Trees

Storing Data



Binary Search Tree

Storing Data



□ ... free position

Binary Search Trees

Probabilistic Model:

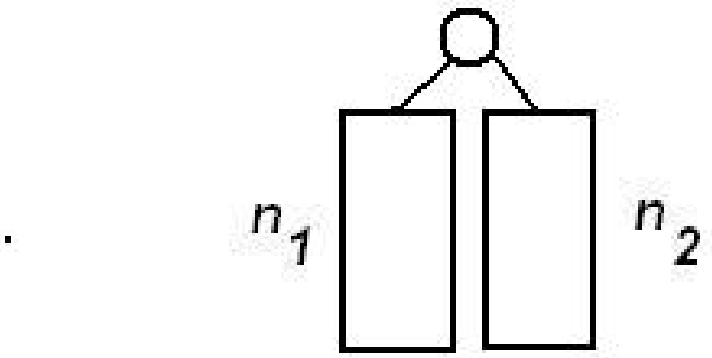
Every permutation of $\{1, 2, \ldots, n\}$ is equally likely.

$\longrightarrow$ probability distribution on binary trees of size n

Binary Search Trees

Height of median of $(2t+1)$**-variant**:
($t = 0$: usual binary search trees)

$$\mathbf{Pr}\{H_n \le k+1\} = \sum_{n_1+n_2=n-1} \frac{\binom{n_1}{t}\binom{n_2}{t}}{\binom{n}{2t+1}} \mathbf{Pr}\{H_{n_1} \le k\} \cdot \mathbf{Pr}\{H_{n_2} \le k\}$$



Binary Search Trees

Generating functions for the median of $2t+1$-variant:

$$\boxed{y_k(x) = \sum_{n \geq 0} \mathbf{Pr}\{H_n \leq k\} \cdot x^n}$$

$$\boxed{y_{k+1}(x)^{(2t+1)} = \frac{(2t+1)!}{(t!)^2} \left(y_k(x)^{(t)}\right)^2}$$

with initial conditions $y_0(x) = 1,\ y_k(0) = 1$.

m-Ary Search Trees

Similarly defined as binary search trees ($m = 2$).

Here every node can store up to $m - 1$ items and has (at most) m subtrees.

There is also a *fringe balanced version* that corresponds to the *median of* $2t + 1$*-variant* in the binary case.

m-Ary Search Trees

$\mathbf{V} = (V_1, V_2 \ldots, V_m)$ ... random vector supported on simplex
$\Delta = \{(s_1, \ldots, s_m) : s_j \geq 0, s_1 + \cdots + s_m = 1\}$ with density

$$f(s_1, \ldots, s_m) = \frac{((t+1)m-1)!}{(t!)^m} (s_1 \cdots s_m)^t.$$

Lemma The functional equation

$$\boxed{F(x/\rho_1) = \mathbb{E}\,(F(xV_1) \cdots F(xV_m))}$$

has (up to scaling) a unique solution $F^{(m,t)}(x)$ with the properties

$$1 - F^{(m,t)}(x) \sim d_1 x^{\beta_1} \log x \quad (x \to 0+)$$

and

$$\lim_{x \to \infty} F^{(m,t)}(x) = 0.$$

m-Ary Search Trees

Let $\beta_1 > 0$ be the positive solution of the equation

$$\sum_{j=0}^{(m-1)(t+1)-1} \log(\beta + t + 1 + j) - \log\left(\frac{(m(t+1))!}{(t+1)!}\right)$$
$$= \sum_{j=0}^{(m-1)(t+1)-1} \frac{\beta}{\beta + t + 1 + j}$$

and set

$$\boxed{\rho_1 = \exp\left(\sum_{j=0}^{(m-1)(t+1)-1} \frac{1}{\beta_1 + t + 1 + j}\right).}$$

m-Ary Search Trees

Theorem (Chauvin and D.)

Let $m \geq 2$ and $t \geq 0$ be integers. There exist sequences c_k with

$$\lim_{k\to\infty} \frac{c_{k+1}}{c_k} = \rho_1$$

such that

$$\boxed{\mathbb{P}\{H_n^{(m,t)} \leq k\} = F^{(m,t)}(n/c_k) + o(1).}$$

Futher there exists $\eta > 0$ with

$$\boxed{\mathbb{P}\{|H_n^{(m,t)} - \mathbb{E}\, H_n^{(m,t)}| \geq y\} = O(e^{-\eta y}).}$$

In particular we have, as $n \to \infty$,

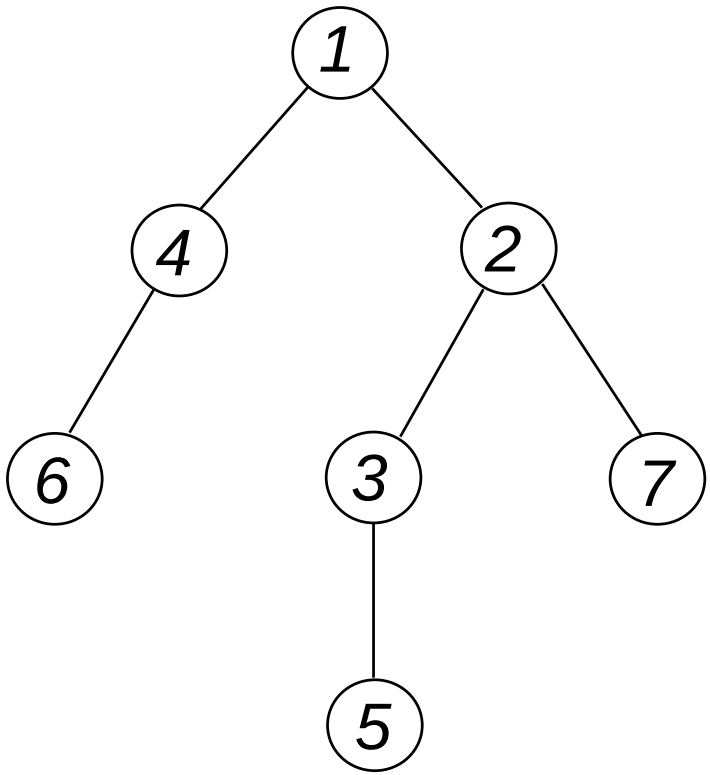
$$\mathbb{V}H_n^{(m,t)} = O(1).$$

Recursive Trees

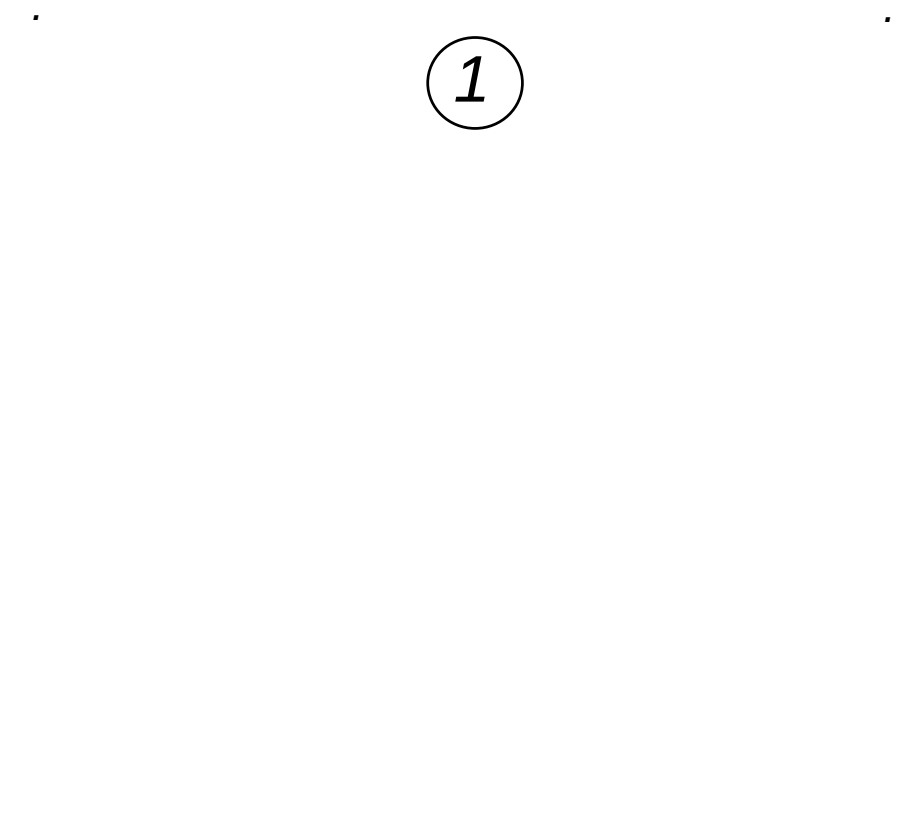
Combinatorial Description:

- labeled rooted tree
- labels are strictly increasing (starting at the root)
- no left-to-right order (non-planar)

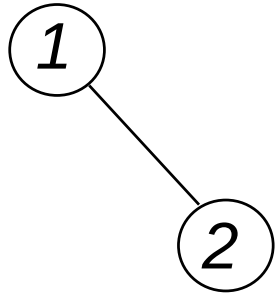
Recursive Trees



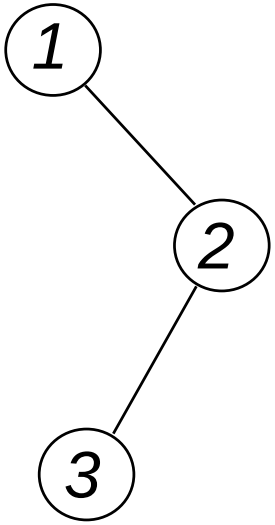
Recursive Trees



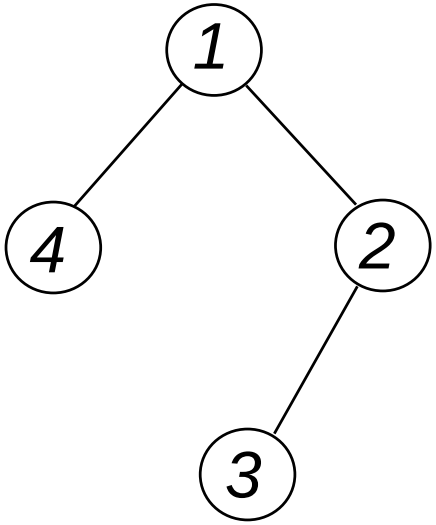
Recursive Trees



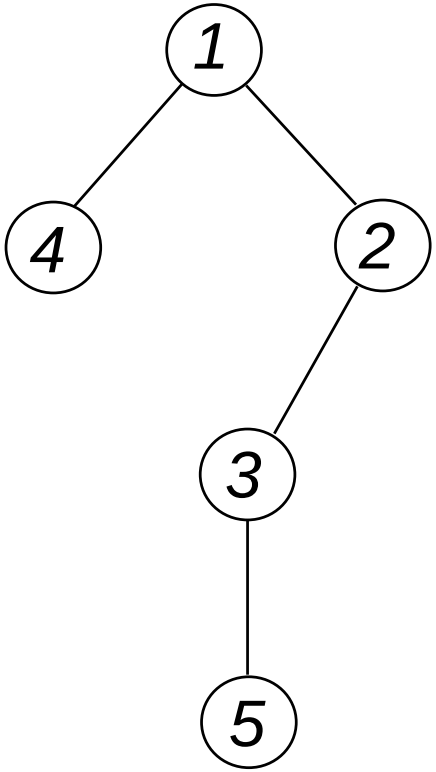
Recursive Trees



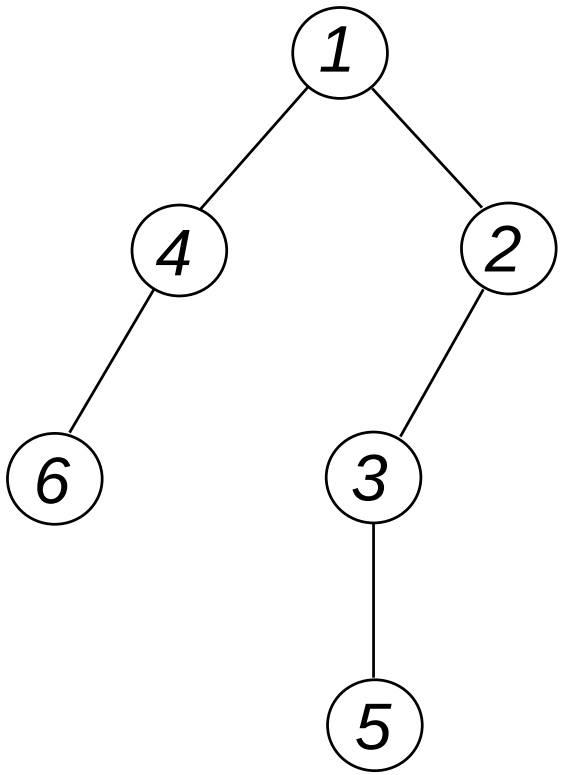
Recursive Trees



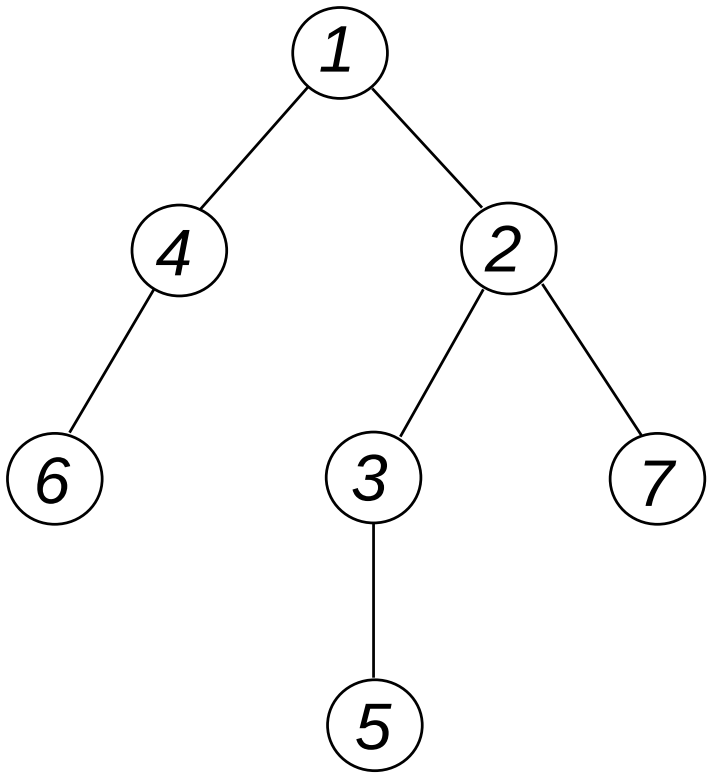
Recursive Trees



Recursive Trees

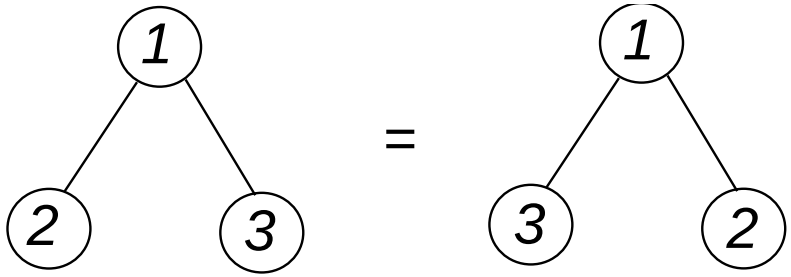


Recursive Trees



Recursive Trees

Remark



Recursive Trees

Number of Recursive Trees:

$$\begin{aligned} y_n &= \text{number of recusive trees of size } n \\ &= (n-1)! \end{aligned}$$

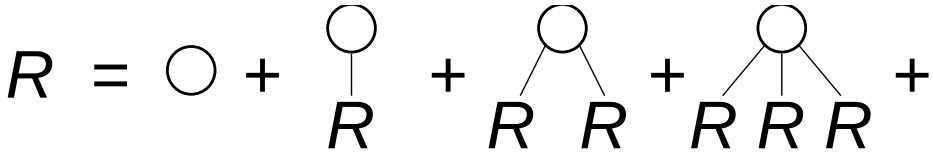
The node with label j has exactly $j-1$ possibilities to be inserted $\Longrightarrow y_n = 1 \cdot 2 \cdots (n-1)$.

Recursive Trees

Generating Functions:

$$y(x) = \sum_{n\geq 1} y_n \frac{x^n}{n!} = \sum_{n\geq 1} \frac{x^n}{n} = \log \frac{1}{1-x}$$

$$\boxed{y'(x) = 1 + y(x) + \frac{y(x)^2}{2!} + \frac{y(x)^3}{3!} + \cdots = e^{y(x)}}$$



A recursive tree can be interpreted as a root followed by an unordered sequence of recursive trees. ($y'(x) = \sum_{n\geq 0} y_{n+1} x^n/n!$)

Recursive Trees

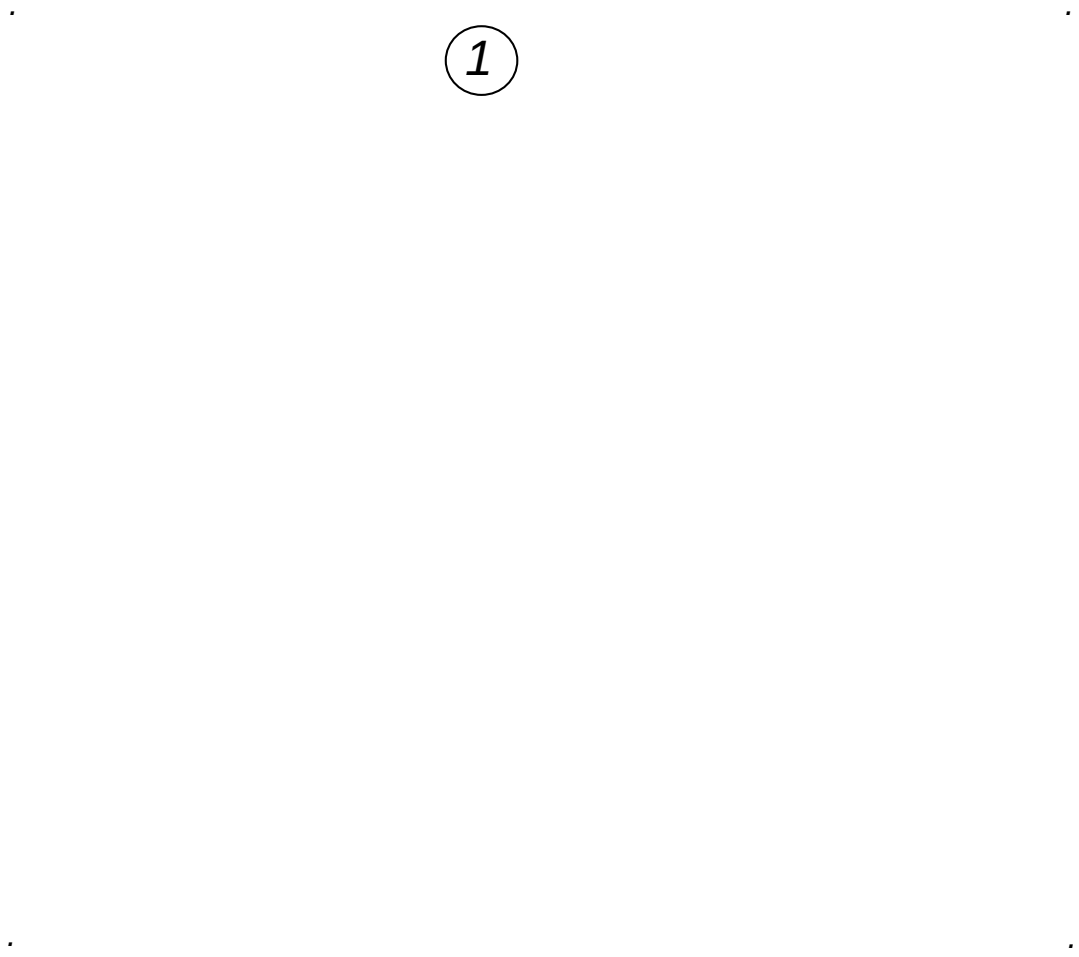
Probability Model:

Process of growing trees

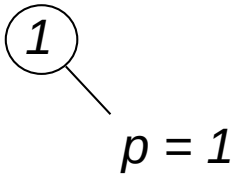
- The process starts with the root that is labeled with 1.
- At step j a new node (with label j) is attached to any previous node with probability $1/(j-1)$.

After n steps every tree (of size n) has equal probability $1/(n-1)!$.

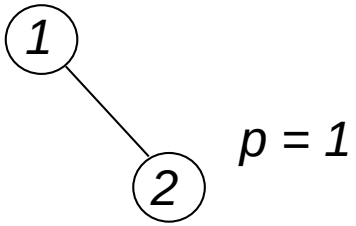
Recursive Trees



Recursive Trees

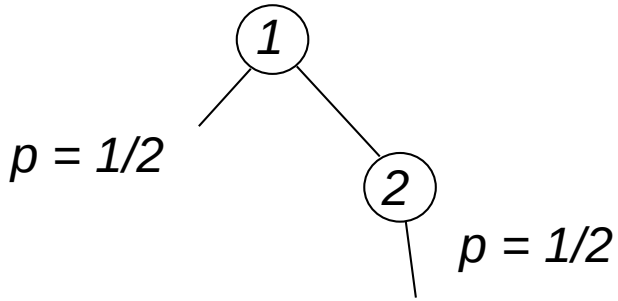


Recursive Trees

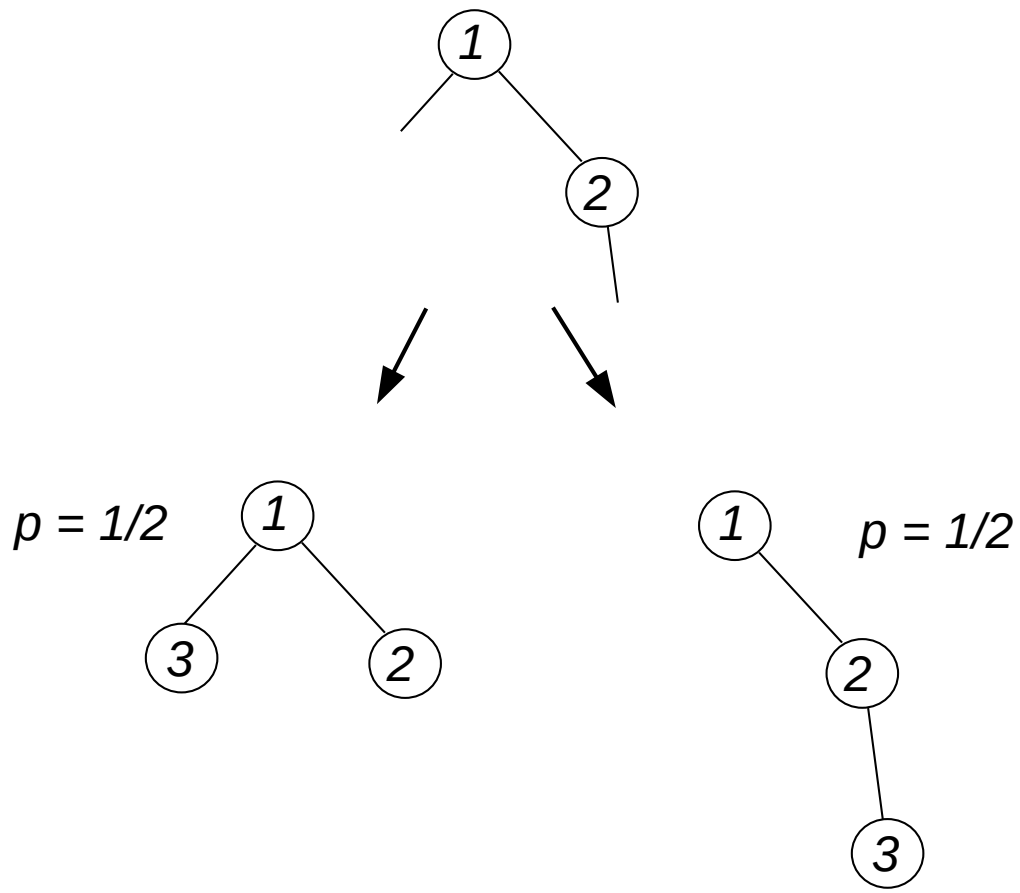


p = 1

Recursive Trees



Recursive Trees



Recursive Trees

Theorem

Let H_n denote the height of random recursive trees of size n.

$$\mathbb{E}\, H_n = e \log n + O\Big(\sqrt{\log n}\,(\log\log n)\Big).$$

Furthermore we have (uniformly for all $k \geq 0$ as $n \to \infty$)

$$\boxed{\mathbb{P}\{H_n \leq k\} = F(n/y'_k(1)) + o(1),}$$

where $F(y)$ satisfies the integral equation

$$y\, F(y/e^{1/e}) = \int_0^y F(z/e^{1/e}) F(y-z)\, dz. \tag{1}$$

Moreover, as $n \to \infty$,

$$\mathbb{V}H_n = O(1)$$

and there exist $\eta > 0$ and $c > 0$ such that

$$\boxed{\mathbb{P}\{|H_n - \mathbb{E}\, H_n| \geq y\} \leq c\, e^{-\eta y}}$$

for all $y \geq 0$.

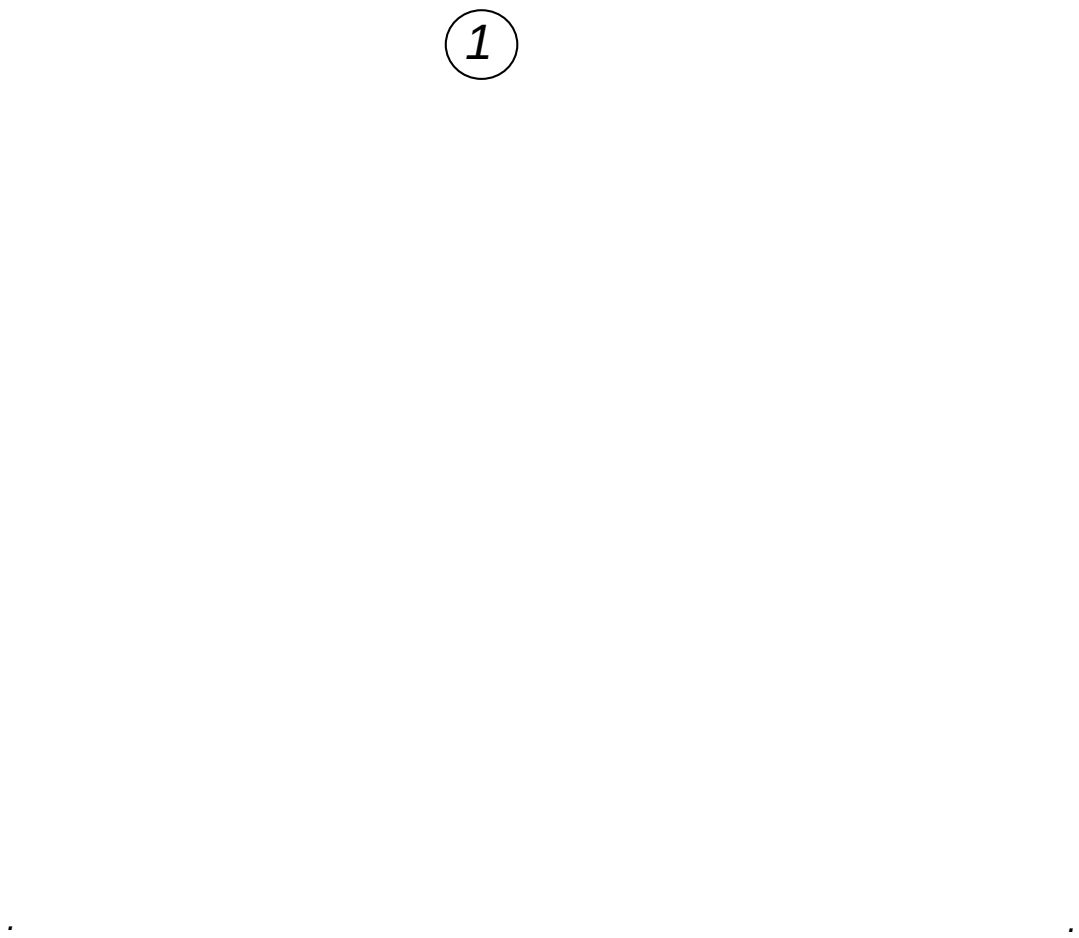
Recursive Trees

Theorem (Szymanski, Devroye and Lu, Goh and Schmutz)

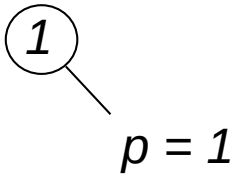
Let $\Delta(T_n)$ denote the maximum out-degree or random recursive trees. Then we have $\mathbb{E}\,\Delta(T_n) \sim \log_2 n$ and the distribution is given by

$$\boxed{\mathbb{P}\{\Delta(T_n) \le k\} = \exp\left(-2^{-(k-\log_2 n+1)}\right) + o(1).}$$

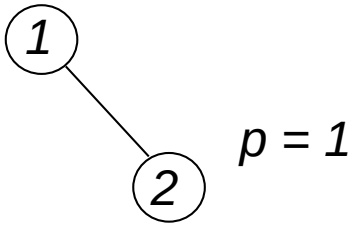
Plane Oriented Trees



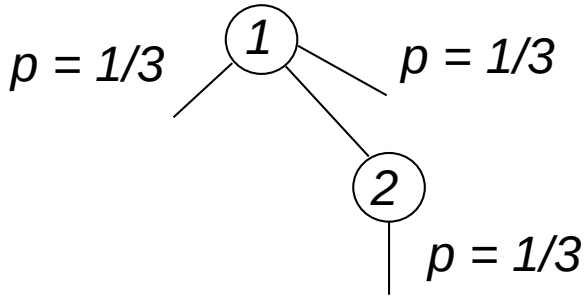
Plane Oriented Trees



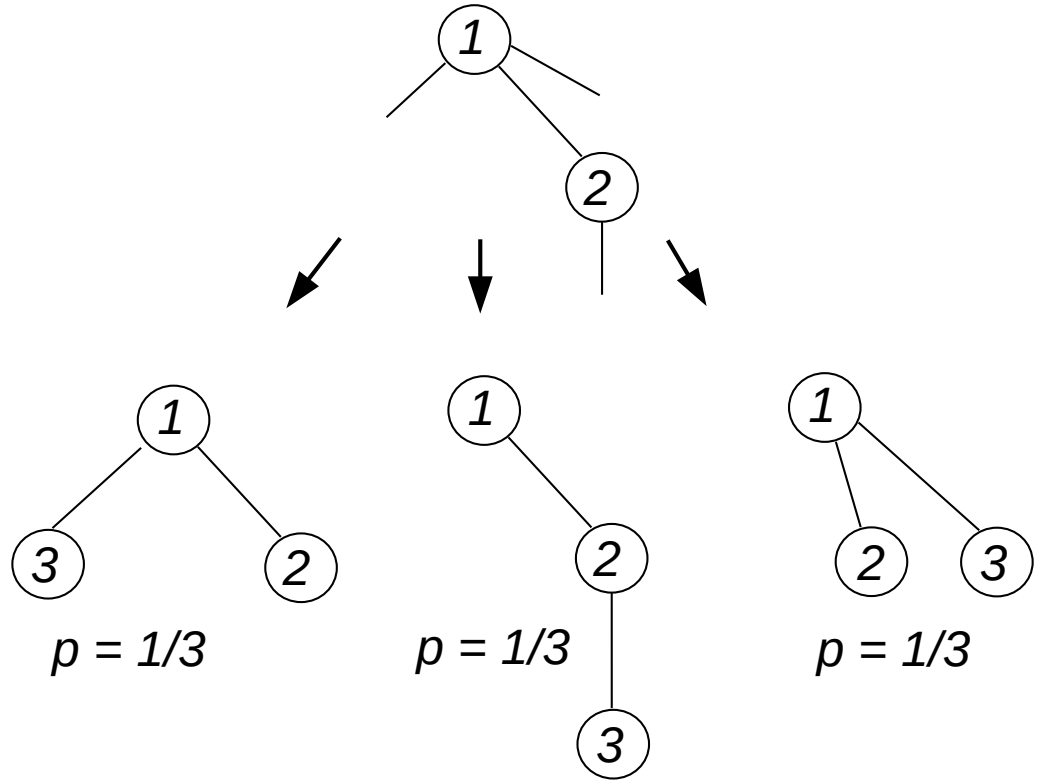
Plane Oriented Trees



Plane Oriented Trees

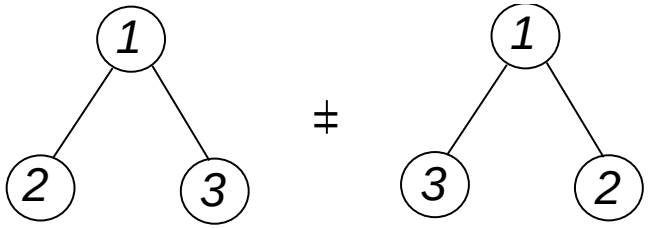


Plane Oriented Trees



Plane Oriented Trees

Remark



Plane Oriented Trees

Number of Plane Oriented Trees:

$$\begin{aligned} y_n &= \text{number of plane oriented trees of size } n \\ &= 1 \cdot 3 \cdot 5 \cdots (2n-3) = (2n-3)!! \\ &= \frac{(2n-2)!}{2^{n-1}(n-1)!} \end{aligned}$$

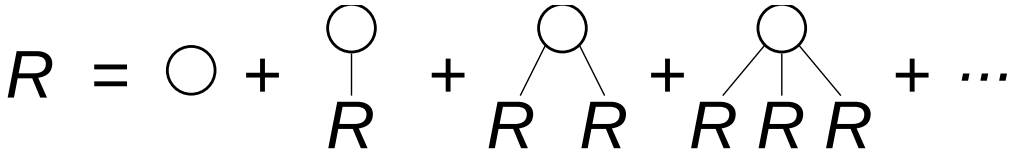
The node with label j has exactly $2j-3$ possibilities to be inserted $\Longrightarrow y_n = 1 \cdot 3 \cdots (2n-3)$.

Plane Oriented Trees

Generating Functions:

$$y(x) = \sum_{n\geq 1} y_n \frac{x^n}{n!} = \sum_{n\geq 1} \frac{1}{2^{n-1}} \binom{2(n-1)}{n-1} \frac{x^n}{n} = 1 - \sqrt{1-2x}$$

$$\boxed{y'(x) = 1 + y(x) + y(x)^2 + y(x)^3 + \cdots = \frac{1}{1-y(x)}}$$



A plane oriented tree can be interpreted as a root followed by an **ordered** sequence of plane oriented trees. ($y'(x) = \sum_{n\geq 0} y_{n+1} x^n/n!$)

Plane Oriented Trees

Probability Model:

Process of growing trees

- The process starts with the root that is labeled with 1.
- At step j a new node (with label j) is attached to any previous node of outdegree d with probability $(d+1)/(2j-3)$.

After n steps every tree (of size n) has equal probability $1/(2n-3)$!!.

Scale Free Trees

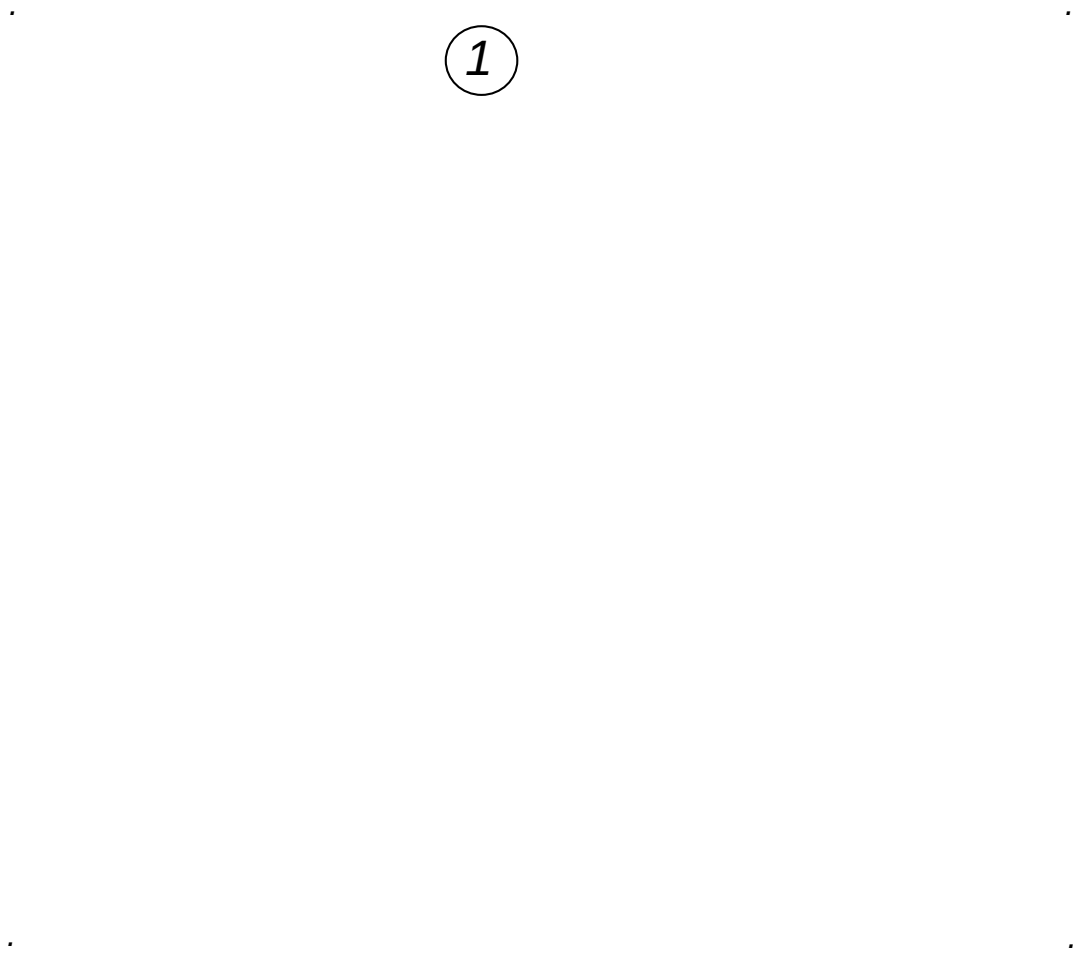
Probability Model:

Process of growing trees

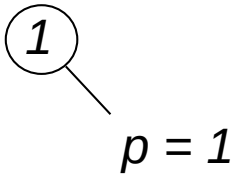
- The process starts with the root that is labeled with 1.
- At step j a new node (with label j) is attached to any previous node of outdegree d with probability proportial to $\boxed{d + r}$ (for some $r > 0$).

For $d = 1$ we get plane oriented trees.

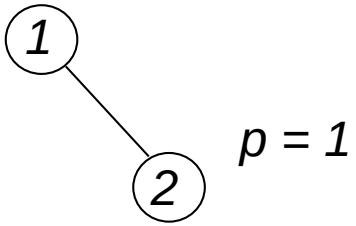
Scale-Free Trees



Scale-Free Trees

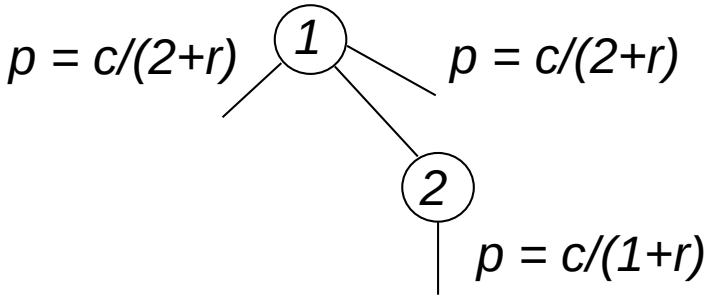


Scale-Free Trees

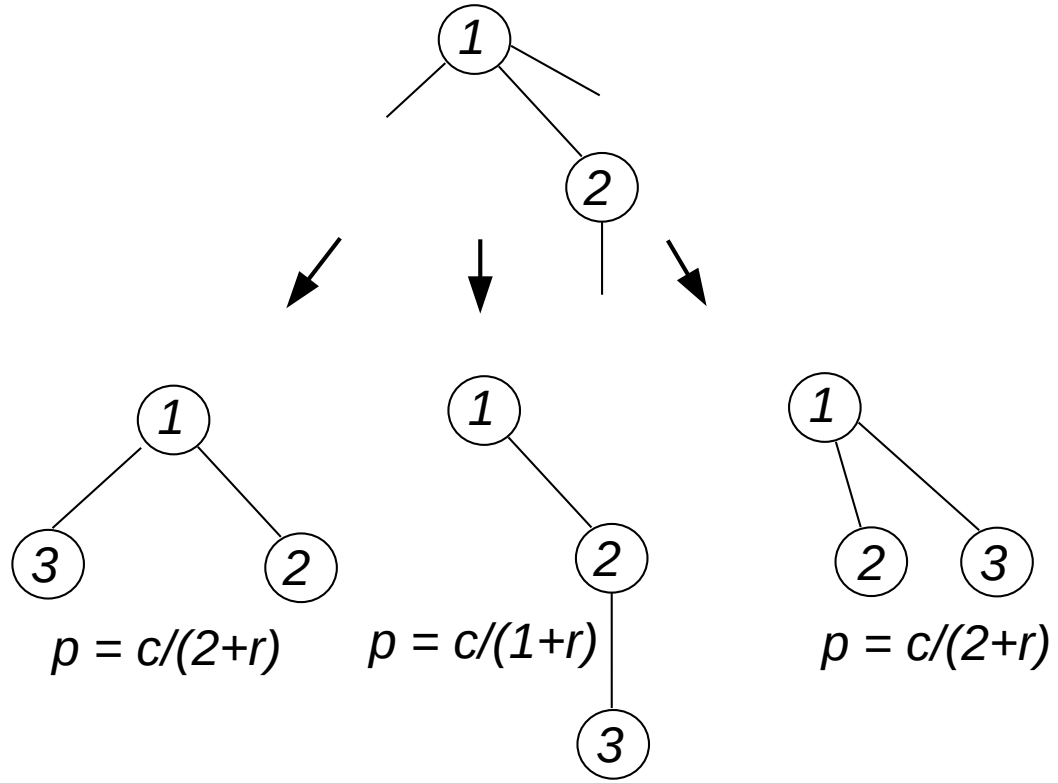


$p = 1$

Scale-Free Trees



Scale-Free Trees



Scale Free Trees

Generating Functions

y_n ... weighted sum of plane oriented trees (according to probability distribution)

$y(x) = \sum_{n \geq 1} y_n \frac{x^n}{n!}$ ... generating function

$$\boxed{y'(x) = \frac{1}{(1 - y(x))^r}}$$

$$\Longrightarrow \quad \boxed{y(x) = 1 - (1 - (r+1)x)^{\frac{1}{r+1}}}$$

Scale Free Trees

Degree distribution

Set

$$\begin{aligned}\lambda_d &= \lim_{n\to\infty} \mathbf{P}\,(\text{a random node in a tree of size } n \text{ has out-degree } d)\\ &= \lim_{n\to\infty} \frac{\text{expected number of nodes with out-degree } d}{n}\end{aligned}$$

Then

$$\boxed{\lambda_d = \frac{(r+1)\Gamma(2r+1)\Gamma(r+d)}{\Gamma(r)\Gamma(2r+d+2)}}$$

We have a **scalefree distribution**

$$\lambda_d \sim \frac{(r+1)\Gamma(2r+1)}{\Gamma(r)} \cdot d^{-2-r}.$$

Scale free trees

Theorem

- Suppose that $r = \frac{A}{B} > 0$ is rational. Then

$$\boxed{\mathbf{P}\{H_n \leq k\} = F(c_k - d_r \log n) + o(1),}$$

where $c_k = k + O(\log k)$ and Set $d_r = 1/((r+1)s)$ with $r\, s\, e^{s+1} = 1$.

Further, $F(x) = \Psi(e^{-x})$, where $\Psi(y)$ is calculated by the following procedure.

Scale free trees

Let $\Phi(y)$ be the solution of

$$y^{\frac{1}{A+B}}\Phi(ye^{-1/d_r}) = \frac{\Gamma\left(1+\frac{1}{A+B}\right)}{\Gamma\left(\frac{1}{A+B}\right)^{A+B+1}}\times$$
$$\times \int_{y_1+\cdots+y_{A+B+1}=y, y_j\geq 0} \prod_{j=1}^{B+1}\left(\Phi(y_j e^{-1/d_r}) y_j^{\frac{1}{A+B}-1}\right)$$
$$\times \prod_{\ell=B+2}^{A+B+1}\left(\Phi(y_\ell) y_\ell^{\frac{1}{A+B}-1}\right) d\mathbf{y}$$

Then

$$\Psi(y) = \frac{\Gamma\left(\frac{A}{A+B}\right)}{\Gamma\left(\frac{1}{A+B}\right)^{A}} \int_{z_1+\cdots+z_A=1, z_j\geq 0} \prod_{j=1}^{A}\left(\Phi(yz_j) z_j^{\frac{1}{A+B}-1}\right) d\mathbf{z}$$

Scale free trees

Further, for all $r > 0$ there exist $\eta > 0$ and $c > 0$ such that

$$\boxed{\mathbb{P}\{|H_n - \mathbb{E}\, H_n| \geq y\} \leq c\, e^{-\eta y}}$$

for all $y \geq 0$.

Remark. The concentration property of the height will be proved in the last part of the talk.

Scale free trees

Theorem (Mori)

Let $\Delta(T_n)$ denote the maximum degree of scale-free trees with parameter $r > 0$. Then

$$\boxed{\frac{\Delta(T_n)}{n^{\frac{1}{1+r}}} \to \mu \quad (a.s.)}$$

for some random variable μ (that is related to the degree distribution of T_n). Further

$$\boxed{\frac{\Delta(T_n) - \mu n^{\frac{1}{1+r}}}{\sqrt{\mu n^{\frac{1}{1+r}}}} \xrightarrow{\mathsf{d}} N(0,1).}$$

Tries

Tries are rooted trees which are used to store data which are labeled with a (possibly) infinite string of symbols from a binary (or generally finite) alphabet.

Each string $\mathbf{x}_j$ defines an infinite path in the (infinite) binary tree. Let u_j denote the node where the suffix part starts. Then we can *trim* the tree by cutting away everything below node u_j. The node u_j becomes now a leaf representing $\mathbf{x}_j$. If we repeat this procedure for all labels $\mathbf{x}_1, \ldots, \mathbf{x}_n$ then we obtain a finite binary tree with n nodes, called the **trie**.

It is usual to assume that the strings follow the Bernoulli model with probability $p \in (0, 1)$, or equivalently they are obtained from a memoryless source on 2 symbols.

Tries

Theorem (Flajolet)

Let H_n denote the height of tries in the symmetric case $p = \frac{1}{2}$. Then

$$\boxed{\mathbb{P}\{H_n \le k\} = \exp\left(-2^{-(k-2\log_2 n)-1}\right) + o(1).}$$

Furthermore, as $n \to \infty$,

$$\mathbb{E}\, H_n = 2\log_2 n + O(1), \quad \mathbb{V}H_n = O(1)$$

and there exist $\eta > 0$ and $c > 0$ such that

$$\boxed{\mathbb{P}\{|H_n - \mathbb{E}\, H_n| \ge y\} \ll e^{-\eta\, y}}$$

for all $y \ge 0$.

Remark. $\mathbb{P}\{H_n \le k\} = \frac{n!}{2^{kn}}\binom{2^k}{n}$

Tries

Theorem (Devroye)

Let H_n denote the height of tries of generated by iid labels with $f(x)$ as a density on $[0,1]$. If $C := \int_0^1 f(x)^2\,dx < \infty$ then

$$\boxed{\mathbb{P}\{H_n \le k\} = \exp\left(-C\,2^{-(k-2\log_2 n)-1}\right) + o(1).}$$

Furthermore,

$$\mathbb{E}\,H_n = 2\log_2 n + O(1).$$

If $\int_0^1 f(x)^2\,dx = \infty$ then $\mathbb{E}\,H_n = \infty$ for all $n \ge 2$.

Tries

Theorem (Pittel)

Let H_n denote height of tries of generated by iid labels on an m-ary alphabet with probability distribution $p_1, p_2, \ldots, p_m$ and set

$$b = \left(\sum_{i=1}^{m} p_i^2 \right)^{-\frac{1}{2}}.$$

Then

$$\boxed{\mathbb{P}\{H_n \leq k\} = \exp\left(-\frac{1}{2} b^{-2(k - \log_b n)} \right) + o(1).}$$

Digital Search Trees

Digital search trees are again rooted trees which are used to store data which are labeled with a (possibly) infinite string of symbols from a binary (or generally finite) alphabet.

The empty string is stored in the root, while the first item occupies the right or left child of the root depending whether its first symbol is "1" or "0". The remaining items are always stored in the next available node according to the rule that we move to the right if the next symbol is "1" and we move to the left if the next symbol is "0". In the same way we can also search for a specific item.

We assume that the 0-1-strings follow the Bernoulli model with probability $p \in (0, 1)$

Digital Search Trees

Theorem

Let H_n denote the height of digital search trees in the symmetric case $p = \frac{1}{2}$. Then there exists a sequence k_n that is asymptotically given by

$$k_n = \log_2 n + \sqrt{2 \log_2 n} - \log_2 \left(\sqrt{2 \log_2 n} \right) + O(1)$$

such that

$$\boxed{\mathbb{P}\{k_n \leq H_n \leq k_n + 1\} = 1 + o(1).}$$

Furthermore, if $0 \leq y \leq c_1 \sqrt{\log n}$ (for some constant $c_1 > 0$) then there exist constants $c_2, c_3 > 0$ with

$$\boxed{\mathbb{P}\{|H_n - k_n| \geq y\} \leq c_2 \, e^{-y c_3 \sqrt{\log_2 n}}.}$$

Height of Scale-Free Trees

y_n ... weighted sum of plane oriented trees (according to probability distribution)

$y(x) = \sum_{n \geq 1} y_n \frac{x^n}{n!}$ ... generating function

$$\boxed{y'(x) = \frac{1}{(1 - y(x))^r}}$$

$$\Longrightarrow \quad \boxed{y(x) = 1 - (1 - (r+1)x)^{\frac{1}{r+1}}}$$

$$y_n = n!(-1)^{n-1}(r+1)^n \binom{1/(r+1)}{n}.$$

Height of Scale-Free Trees

$$y_k(z) = \sum_{n \geq 0} y_n \mathbb{P}\{H_n \leq k\} \frac{z^n}{n!}$$

$$\Longrightarrow \quad \boxed{y'_{k+1}(z) = \frac{1}{(1 - y_k(z))^r}} \qquad (y_0(z) = 0,\ y_{k+1}(0) = 0).$$

$$Y_k(z) := y'_k(z)$$

$$\Longrightarrow \quad \boxed{Y'_{k+1}(z) = r\, Y_{k+1}(z)^{1+\frac{1}{r}} Y_k(z)} \qquad (Y_1(z) = 1,\ Y_{k+1}(0) = 1).$$

$$Y_k(z) = y'_k(z) = \sum_{n \geq 0} y_{n+1} \mathbb{P}\{H_{n+1} \leq k\} \frac{z^n}{n!}.$$

Height of Scale-Free Trees

Lemma 1

$Y_1(z), Y_2(z), \overline{Y}_1(z), \overline{Y}_2(z)$ ... non-negative, continuous functions, defined for $z \geq 0$

$$Y_1(0) < \overline{Y}_1(0), \quad Y_2(0) < \overline{Y}_2(0),$$

$$Y_2'(z) = r\, Y_2(z)^{1+\frac{1}{r}} Y_1(z), \quad \overline{Y}_2'(z) = r\, \overline{Y}_2(z)^{1+\frac{1}{r}} \overline{Y}_1(z).$$

$$\overline{Y}_1(z) - Y_1(z) \quad \text{has exactly one positive zero}$$

$\Longrightarrow$ $\overline{Y}_2(z) - Y_2(z)$ has at most one positive zero.

Height of Scale-Free Trees

Proof of Lemma 1.

$$y_j(z) := \int_0^z Y_j(t)\,dt \quad \text{and} \quad \overline{y}_j(z) = \int_0^z \overline{Y}_j(t)\,dt \qquad (j = 1, 2)$$

$$\overline{y}_1'(z) - y_1'(z) = \overline{Y}_1(z) - Y_1(z)$$

$\Longrightarrow$ $\overline{y}_1'(z) - y_1'(z)$ has exactly one positive zero ζ.

$\Longrightarrow$ $\overline{y}_1(z) - y_1(z)$ increasing for $0 \le z \le \zeta$ and decreasing for $z \ge \zeta$.

$\Longrightarrow$ $\overline{y}_1(z) - y_1(z)$ has at most one positive zero.

$$\Longrightarrow \quad \overline{Y}_2(z) - Y_2(z) = \overline{y}_2'(z) - y_2'(z) = \frac{(1-\overline{y}_1(z))^{-r} - (1-y_1(z))^{-r}}{\overline{y}_1(z) - y_1(z)}\,(\overline{y}_1(z) - y_1(z))$$

has at most one positive zero, too.

Height of Scale-Free Trees

Lemma 2

The sequence $Y_k(1/(r+1))$ is log-concave, that is,

$$\boxed{\frac{Y_{k+2}(1/(r+1))}{Y_{k+1}(1/(r+1))} \leq \frac{Y_{k+1}(1/(r+1))}{Y_k(1/(r+1))}.}$$

Corollary

The limit

$$\lim_{k\to\infty} \frac{Y_{k+1}(1/(r+1))}{Y_k(1/(r+1))} =: \rho \geq 1$$

exists.

Height of Scale-Free Trees

Proof of Lemma 2

For $0 \le \gamma < 1$ set

$$V_k(z,\gamma) = \begin{cases} (1-(r+1)z)^{-r/(r+1)} & \text{for } 0 \le z \le \frac{1}{r+1}(1-\gamma), \\ \gamma^{-r/(r+1)} Y_k\left(\frac{z-\frac{1}{r+1}(1-\gamma)}{\gamma}\right) & \text{for } \frac{1}{r+1}(1-\gamma) \le z \le \frac{1}{r+1}. \end{cases}$$

$$\Longrightarrow \quad \boxed{V'_{k+1}(z,\gamma) = r\, V_{k+1}(z,\gamma)^{1+\frac{1}{r}} V_k(z,\gamma),}$$

($V_k(0) = 1$ and $V_k(1/(r+1),\gamma) = \gamma^{-r/(r+1)} Y_k(1/(r+1))$.)

Lemma 1 $\Longrightarrow$ $Y_{k+1}(z) - V_k(z,\gamma)$ has (at most) one positive zero.

Height of Scale-Free Trees

$$\gamma_k := \left(\frac{Y_k(1/(r+1))}{Y_{k+1}(1/(r+1))} \right)^{1+\frac{1}{r}}$$

$$\Longrightarrow \quad V_k(1/(r+1), \gamma_k) = Y_{k+1}(1/(r+1)).$$

$$\Longrightarrow \quad \zeta = 1/(r+1) \text{ is the only positive zero of } Y_{k+1}(z) - V_k(z, \gamma).$$

$$\Longrightarrow \quad Y_{k+1}(z) \leq V_k(z, \gamma_k) \quad \text{for} \quad 0 \leq z \leq \frac{1}{r+1}$$

+ integration

$$\begin{aligned} \Longrightarrow \quad Y_{k+2}(1/(r+1)) &\leq V_{k+1}(1/(r+1), \gamma_k) \\ &= \gamma_k^{-r/(r+1)} Y_{k+1}(1/(r+1)) \\ &= \frac{Y_{k+1}(1/(r+1))^2}{Y_k(1/(r+1))}. \end{aligned}$$

Height of Scale-Free Trees

Lemma 3

For $0 \leq z < \frac{1}{r+1}$ and $k \geq 1$ we have

$$Y(z) - Y_k(z) \leq \left(\frac{2r+1}{r+1}\right)^k \sum_{\ell \geq k} \frac{1}{\ell!} \left(\log \frac{1}{1-(r+1)z)}\right)^\ell.$$

Height of Scale-Free Trees

Proof of Lemma 3

By induction:

$k = 1$: $Y_1(z) = 1$ and $(1-(r+1)z)^{-r/(r+1)} \leq (1-(r+1)z)^{-1}$.

$k \to k+1$:

$$\begin{aligned} Y(z)' - Y_{k+1}(z)' &= r\left(Y(z)^{2+\frac{1}{r}} - Y_{k+1}(z)^{1+\frac{1}{r}} Y_k(z)\right) \\ &\leq r\left(Y(z)^{2+\frac{1}{r}} - Y_k(z)^{2+\frac{1}{r}}\right) \\ &\leq r\left(2+\frac{1}{r}\right) Y(z)^{1+\frac{1}{r}} \left(Y(z) - Y_k(z)\right) \\ &= (2r+1)\frac{1}{1-(r+1)z}\left(Y(z) - Y_k(z)\right) \end{aligned}$$

Height of Scale-Free Trees

Lemma 4

Suppose that $1 < C < e^{r/(r+1)}$ and $C > \left(2 + \frac{1}{r}\right) e \log C$

$$\Longrightarrow \quad \frac{Y_{k+1}(1/(r+1))}{Y_k(1/(r+1))} \geq C$$

for all $k \geq 1$.

Corollary. $\boxed{\rho > 1}$.

Height of Scale-Free Trees

Proof of Lemma 4. Set $c = C^{-r/(r+1)}$ and $z_0 = \frac{1}{r+1}(1 - c^k)$.

Lemma 3 $\Longrightarrow$ (with some $c_1 > 0$)

$$\begin{aligned} Y(z_0) - Y_k(z_0) &\le \left(\frac{2r+1}{r+1}\right)^k \sum_{\ell \ge k} \frac{1}{\ell!} \left(k \log \frac{1}{c}\right)^\ell \\ &\le c_1 \left(\frac{2r+1}{r+1}\right)^k \frac{\left(k \log \frac{1}{c}\right)^k}{k!} \\ &\le c_1 \left(\frac{2r+1}{r+1} e \log \frac{1}{c}\right)^k \\ &= c_1 \left(\left(2 + \frac{1}{r}\right) e \log C\right)^k. \end{aligned}$$

$Y(z_0) = C^k$, $C > \left(2 + \frac{1}{r}\right) e \log C$

$$\Longrightarrow \quad Y_k(1/(r+1)) \ge Y_k(z_0) \ge C^k(1 + o(1)).$$

Height of Scale-Free Trees

$$\Longrightarrow \quad \rho = \lim_{k\to\infty} \frac{Y_{k+1}(1/(r+1))}{Y_k(1/(r+1))} \geq C.$$

Lemma 2 $\Longrightarrow$ the sequence $\frac{Y_{k+1}(1/(r+1))}{Y_k(1/(r+1))}$ is decreasing.

$$\Longrightarrow \quad \frac{Y_{k+1}(1/(r+1))}{Y_k(1/(r+1))} \geq C.$$

Height of Scale-Free Trees

Lemma 5

If $n \geq Y_k(1/(r+1))^{1+\frac{1}{r}}$ then

$$\boxed{\mathbb{P}\{H_n \leq k\} = O\left(Y_k(1/(r+1)) \cdot n^{-r/(r+1)}\right).}$$

Conversely if $n \leq Y_k(1/(r+1))^{1+\frac{1}{r}}$ then

$$\boxed{\mathbb{P}\{H_n > k\} = O\left(Y_k(1/(r+1))^{-1-\frac{1}{r}} \cdot n\right).}$$

Height of Scale-Free Trees

Proof of Lemma 5.

Define z_k by $\boxed{Y(z_k) = Y_k(1/(1+r))}$.

$$\Longrightarrow \quad z_k = \frac{1}{r+1}\left(1 - \eta_k^{-1-\frac{1}{r}}\right) \quad \text{with} \quad \eta_k = Y_k(1/(1+r)).$$

Set

$$\tilde{Y}(z) = (z_k(r+1))^{r/(r+1)} Y(z_k(r+1)z)$$

$$\Longrightarrow \boxed{\tilde{Y}'(z) = r\tilde{Y}(z)^{2+\frac{1}{r}}.}$$

$\tilde{Y}(0) < 1$ + Lemma 1 $\Longrightarrow$

- $\tilde{Y}(z) \leq Y_k(z)$ for $0 \leq z \leq z_k$,
- $\tilde{Y}(z) \geq Y_k(z)$ for $z \geq z_k$.

Height of Scale-Free Trees

$\mathbb{P}\{H_{n+1} \le k\} \le \mathbb{P}\{H_n \le k\}$

$z \ge z_k \Longrightarrow$

$$\begin{aligned}\tilde{Y}(z) &\ge Y_k(z)\\ &\ge \sum_{\ell=0}^{n-1} y_{\ell+1}\mathbb{P}\{H_{\ell+1} \le k\}\frac{z^\ell}{\ell!}\\ &\ge \mathbb{P}\{H_n \le k\}\sum_{\ell=0}^{n-1} y_{\ell+1}\frac{z^\ell}{\ell!}\end{aligned}$$

$n \ge \eta_k^{1+\frac{1}{r}}$, $z = \frac{1}{r+1}$:

$$\tilde{Y}(1/(r+1)) \le c_2\eta_k,$$

$$\sum_{\ell=0}^{n-1} y_{\ell+1}\frac{(r+1)^\ell}{\ell!} \ge c_2 n^{r/(r+1)}$$

$$\Longrightarrow \quad \boxed{\mathbb{P}\{H_n \le k\} \le c_4\eta_k n^{-r/(r+1)}.}$$

Height of Scale-Free Trees

$0 \le z \le z_k \implies$

$$\begin{aligned} Y(z) - \tilde{Y}(z) &\ge Y(z) - Y_k(z) \\ &\ge \sum_{\ell=n-1}^{\infty} y_{\ell+1} \mathbb{P}\{H_{\ell+1} > k\} \frac{z^\ell}{\ell!} \\ &\ge \mathbb{P}\{H_n > k\} \sum_{\ell=n-1}^{\infty} y_{\ell+1} \frac{z^\ell}{\ell!}. \end{aligned}$$

$n \le \eta_k^{1+\frac{1}{r}}$, $z' = \frac{1}{r+1}\left(1 - \frac{1}{n}\right) \le z_k$.

$$Y(z') - \tilde{Y}(z') \le c_5 n^{1+\frac{r}{r+1}} \eta_k^{-1-\frac{1}{r}},$$

$$\sum_{\ell=n-1}^{\infty} y_{\ell+1} \frac{(z')^\ell}{\ell!} \ge c_6\, n^{r/(r+1)}$$

$$\implies \quad \boxed{\mathbb{P}\{H_n > k\} \le c_7 \eta_k^{-1-\frac{1}{r}}\, n}$$

Height of Scale-Free Trees

Lemma 6

Let $k(n) := \max\{\ell \geq 1 : Y_\ell(1/(r+1))^{1+\frac{1}{r}} \leq n\}$. Then

$$\mathbb{E}\, H_n = k(n) + O(1) \tag{2}$$

and there exist $\eta > 0$ and $c > 0$ such that

$$\boxed{\mathbb{P}\{|H_n - \mathbb{E}\, H_n| > y\} \leq ce^{-\eta y}}$$

for all $y > 0$.

Proof of Lemma 6.

Lemma 4 $\Longrightarrow$

$$\frac{Y_{k(n)+\ell}(1/(r+1))}{Y_{k(n)}(1/(r+1))} \geq C^\ell \qquad (\ell \geq 0).$$

$$\begin{aligned}\Longrightarrow \quad \mathbb{P}\{H_n > k(n)+\ell\} &\le c_8 Y_{k(n)+\ell}(1/(r+1))^{-1-\frac{1}{r}}\, n \\ &\le c_8\, C^{\frac{r+1}{r}\ell}\, Y_{k(n)}(1/(r+1))^{-1-\frac{1}{r}}\, n \\ &\le c_9 C^{-\frac{r+1}{r}\ell}.\end{aligned}$$

Similarly

$$\mathbb{P}\{H_n \le k(n)-\ell\} \le c_{10} C^{-\ell}.$$

Height of Scale-Free Trees

Remark

The above proof provides stong concentration around the mean but it does not say where the mean value actually is.

It is related to the the actual growth of $Y_k(1/(r+1))$. Thus one has to analyze the recurrence $Y'_{k+1}(z) = r\Big(Y_{k+1}(z)\Big)^{1+\frac{1}{r}} Y_k(z)$ in more detail.

Thank You!