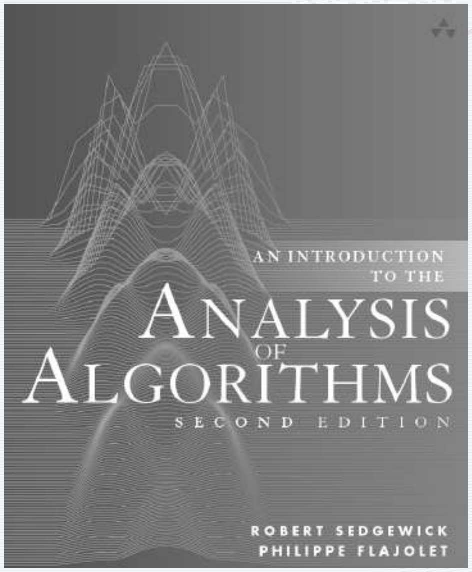


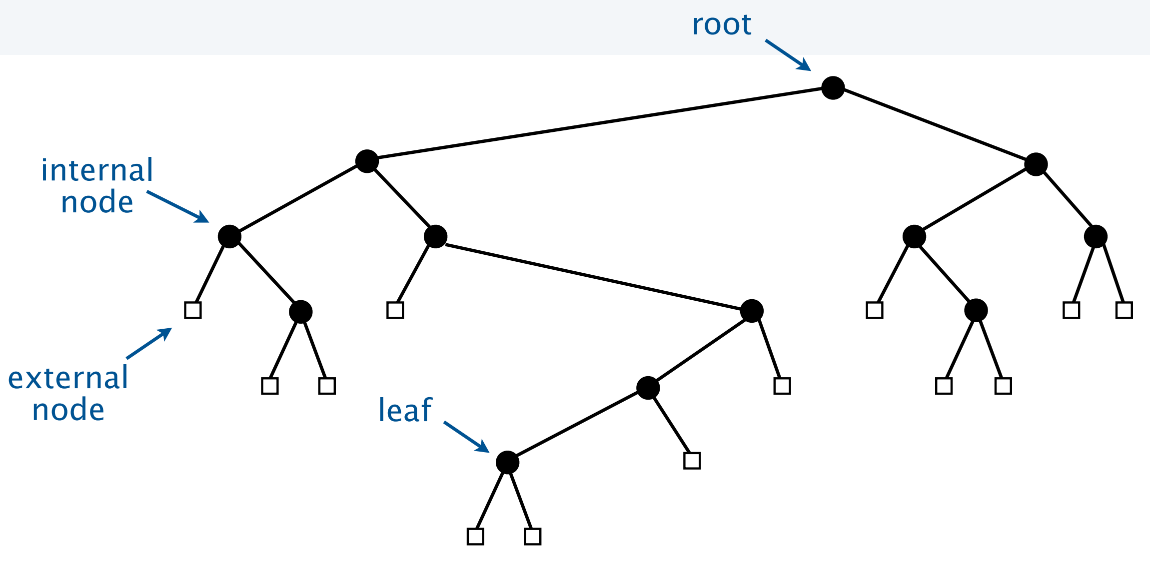
http://aofa.cs.princeton.edu

6. Trees

- Trees and forests
- Binary search trees
- **Path length**
- Other types of trees

Path length in binary trees

Definition. A *binary tree* is an external node or an internal node and two binary trees.



level (depth)	internal path length $ipl(t)$	external path length $xpl(t)$
0	$0 \cdot 1$	$0 \cdot 0$
1	$+\ 1 \cdot 2$	$+\ 1 \cdot 0$
2	$+\ 2 \cdot 4$	$+\ 2 \cdot 0$
3	$+\ 3 \cdot 3$	$+\ 3 \cdot 5$
4	$+\ 4 \cdot 1$	$+\ 4 \cdot 5$
5	$+\ 5 \cdot 1$	$+\ 5 \cdot 1$
6 (height $h(t)$)	$= 28$	$+\ 6 \cdot 2$
		$= 52$

internal path length: $$ipl(t) = \sum_{k \geq 0} k \cdot \{\#\ \text{internal nodes at depth k}\}$$

external path length: $$xpl(t) = \sum_{k \geq 0} k \cdot \{\#\ \text{external nodes at depth k}\}$$

Path length in binary trees

notation	*definition*
t	binary tree
$\|t\|$	# internal nodes in t
$\boxed{t}$	# external nodes in t
t_L and t_R	left and right subtrees of t
$ipl\,(t)$	internal path length of t
$xpl\,(t)$	external path length of t

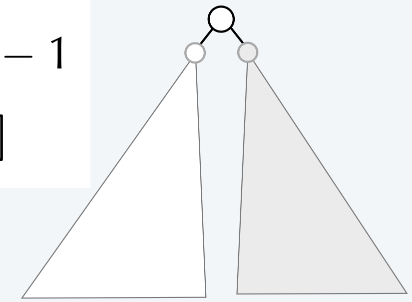
recursive relationships

$$|\,t\,| = |\,t_L\,| + |\,t_R\,| + 1$$

$$\boxed{t} = \boxed{t_L} + \boxed{t_R}$$

$$ipl(t) = ipl(t_L) + ipl(t_R) + |\,t\,| - 1$$

$$xpl(t) = xpl(t_L) + xpl(t_R) + \boxed{t}$$



Lemma 1. $\boxed{t} = |\,t\,| + 1$

Proof. Induction.

$$\boxed{t} = \boxed{t_L} + \boxed{t_R}$$

$$= |\,t_L\,| + 1 + |\,t_R\,| + 1$$

$$= |\,t\,| + 1$$

Lemma 2. $xpl(t) = ipl(t) + 2|\,t\,|$

Proof. Induction.

$$xpl(t) = xpl(t_L) + xpl(t_R) + \boxed{t}$$

$$= ipl(t_L) + 2|\,t_L\,| + ipl(t_R) + 2|\,t_R\,| + |\,t\,| + 1$$

$$= ipl(t) + 2|\,t\,|$$

Problem 1: What is the expected path length of a random binary tree?

Q_{Nk} = # trees with N nodes and ipl k

T_N = # trees

Q_N = cumulated cost (total ipl)

$Q_{10} = 1$

$T_1 = 1$

$Q_1 = 0$

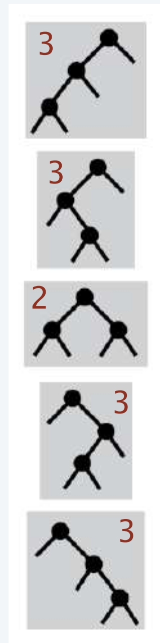
$Q_1/T_1 = 0$

$Q_{21} = 2$

$T_2 = 2$

$Q_2 = 2$

$Q_2/T_2 = 1$



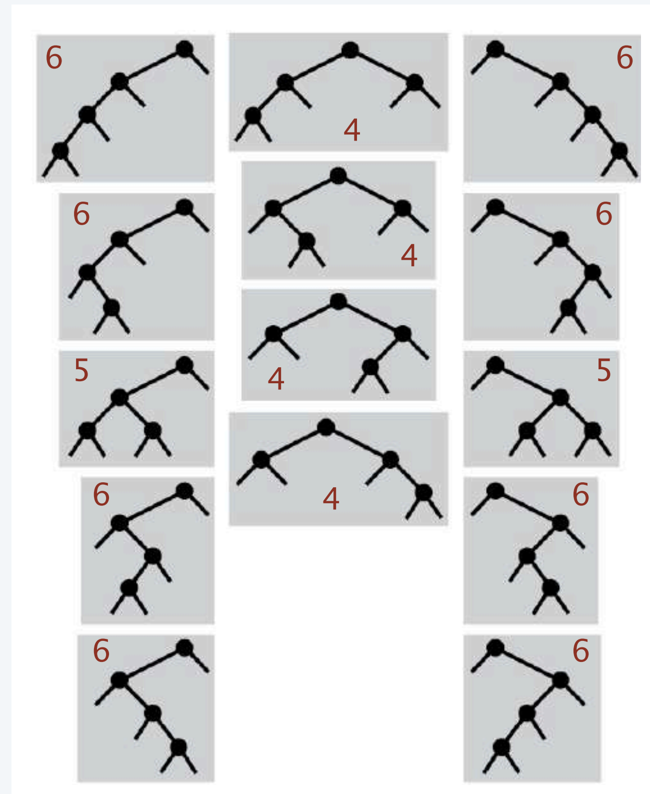
$Q_{32} = 1$

$Q_{33} = 4$

$T_3 = 2$

$Q_3 = 1 \cdot 2 + 4 \cdot 3 = 14$

$Q_3/T_3 = 2.8$



$Q_{44} = 4$

$Q_{45} = 2$

$Q_{46} = 8$

$T_4 = 14$

$Q_4 = 4 \cdot 4 + 2 \cdot 5 + 8 \cdot 6 = 74$

$Q_4/T_4 \doteq 5.286$

Average path length in a random binary tree

T is the set of all binary trees.

$|t|$ is the number of internal nodes in t.

ipl(t) is the internal path length of t.

T_N is the # of binary trees of size N (Catalan).

Q_N is the total ipl of all binary trees of size N.

Counting GF.

$$T(z) = \sum_{t\in T} z^{|t|} = \sum_{N\geq 0} T_N z^N = \sum_{N\geq 0} \frac{1}{N+1}\binom{2N}{N} z^N \sim \frac{4^N}{\sqrt{\pi N^3}}$$

Cumulative cost GF.

$$Q(z) = \sum_{t\in T} \mathrm{ipl}(t) z^{|t|}$$

Average ipl of a random N-node binary tree.

$$\frac{[z^N]Q(z)}{[z^N]T(z)} = \frac{[z^N]Q(z)}{T_N}$$

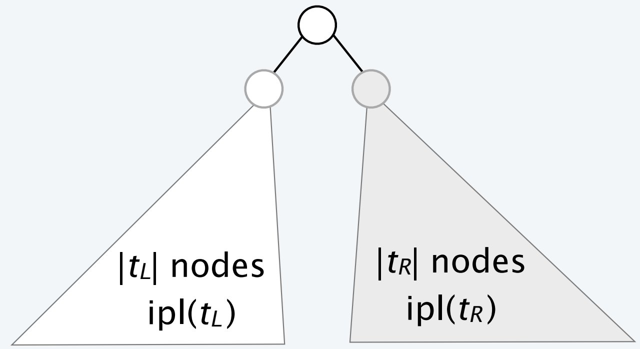
Next: Derive a functional equation for the CGF.

|t_L|

CGF functional equation for path length in binary trees

Counting GF. $$T(z) = \sum_{t \in T} z^{|t|}$$

CGF. $$Q(z) = \sum_{t \in T} ipl(t) z^{|t|}$$



$$ipl(t) = ipl(t_L) + ipl(t_R) + |\, t_L \,| + |\, t_R \,|$$

empty tree

root

Decompose from definition. $$Q(z) = 1 + \sum_{t_L \in T} \sum_{t_R \in T} \big(ipl(t_L) + ipl(t_R) + |t_L| + |t_R|\big) z^{|t_L| + |t_R| + 1}$$

$$\sum_{t_L \in T} ipl(t_L) z^{|t_L|} \sum_{t_R \in T} z^{|t_R|} = Q(z)T(z)$$

$$\sum_{t_L \in T} |t_L| z^{|t_L|} \sum_{t_R \in T} z^{|t_R|} = zT'(z)T(z)$$

$$= 1 + 2zQ(z)T(z) + 2z^2 T'(z)T(z)$$

Expected path length of a random binary tree: full derivation

CGF.
$$Q(z) = \sum_{t \in T} ipl(t) z^{|t|}$$

Decompose from definition.
$$Q(z) = 1 + \sum_{t_L \in T} \sum_{t_R \in T} \big(ipl(t_L) + ipl(t_R) + |t_L| + |t_R|\big) z^{|t_L|+|t_R|+1}$$
$$= 2zT(z)\big(Q(z) + zT'(z)\big)$$

Solve.
$$Q(z) = \frac{2z^2 T(z) T'(z)}{1 - 2zT(z)}$$

Do some algebra (omitted)
$$zQ(z) = \frac{z}{1-4z} - \frac{1-z}{\sqrt{1-4z}} + 1$$

$$T(z) = \frac{1-\sqrt{1-4z}}{2z} \qquad T_N \sim \frac{4^N}{N\sqrt{\pi N}}$$
$$T'(z) = -\frac{1-\sqrt{1-4z}}{2z^2} + \frac{1}{z\sqrt{1-4z}}$$
$$1 - 2zT(z) = \sqrt{1-4z}$$

Expand.
$$Q_N \equiv [z^N]Q(z) \sim 4^N$$

Compute *average internal path length.*
$$Q_N / T_N \sim N\sqrt{\pi N}$$

Problem 2: What is the expected path length of a random *BST*?

C_{Nk} = # *permutations* resulting in a BST with N nodes and ipl k

$N!$ = # permutations

C_N = cumulated cost (total ipl)

$C_{10} = 1$

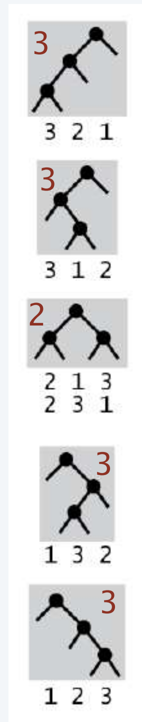
$C_1 = 0$

$C_1/1! = 0$

$C_{21} = 2$

$C_2 = 2$

$C_2/2! = 1$

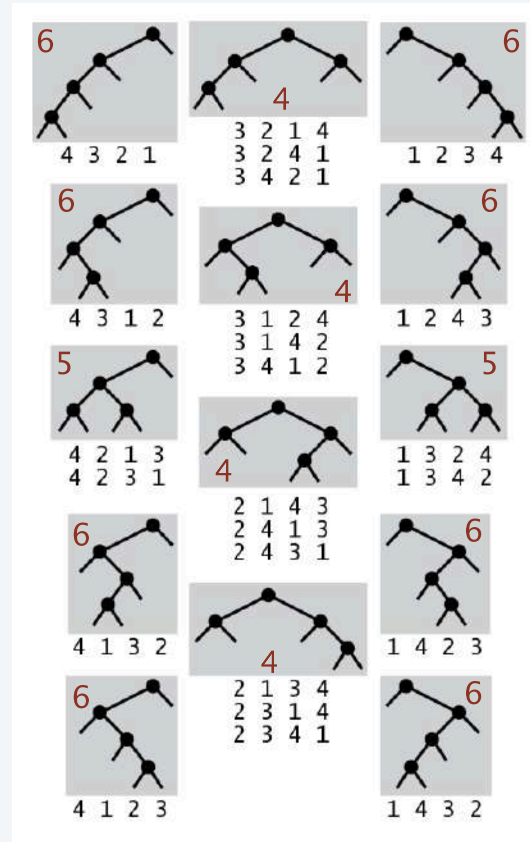


$C_{32} = 2$

$C_{33} = 4$

$C_3 = 2\cdot 2 + 4\cdot 3 = 16$

$C_3/3! \doteq 2.667$



$C_{44} = 12$

$C_{45} = 4$

$C_{46} = 8$

$C_4 = 12\cdot 4 + 4\cdot 5 + 8\cdot 6 = 74$

$C_4/4! \doteq 4.833$

Recall: A property of permutations.

Average path length in a BST built from a random permutation

P is the set of all permutations.

$|p|$ is the length of p.

$\mathrm{ipl}(p)$ is the ipl of the BST built from p by inserting into an initially empty tree.

P_N is the # of permutations of size N ($N!$).

C_N is the total ipl of BSTs built from all permutations.

Counting EGF.
$$P(z) = \sum_{p \in P} \frac{z^{|p|}}{|p|!} = \sum_{N \ge 0} N! \frac{z^N}{N!} = \frac{1}{1-z}$$

Cumulative cost EGF.
$$C(z) = \sum_{p \in P} \mathrm{ipl}(p) \frac{z^{|p|}}{|p|!}$$

Expected ipl of a BST built from a random permutation.
$$\frac{N![z^N]C(z)}{[z^N]P(z)} = \frac{N![z^N]C(z)}{N!} = [z^N]C(z) \longleftarrow$$

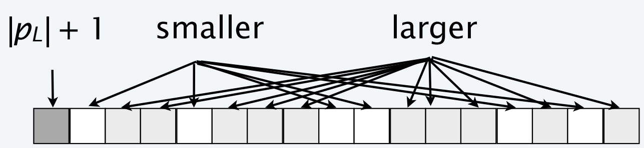
skip a step because counting sequence and EGF normalization are both $N!$

Next: Derive a functional equation for the cumulated cost EGF.

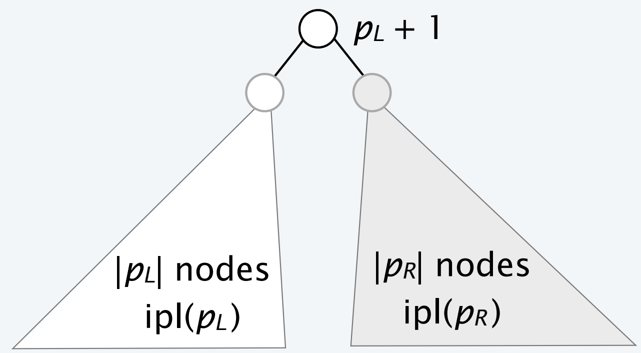
CGF functional equation for path length in BSTs

Cumulative cost EGF. $C(z) = \sum_{p \in P} \text{ipl}(p) \frac{z^{|p|}}{|p|!}$

Counting GF. $P(z) = \sum_{p \in P} \frac{z^{|p|}}{|p|!} = \frac{1}{1-z}$



$\binom{|p_L| + |p_R|}{|p_L|}$ perms lead to the same tree with
- $|p_L|$ +1 at the root
- p_L nodes on the left
- p_R nodes on the right



Decompose. $C(z) = \sum_{p_L \in \mathcal{P}} \sum_{p_R \in \mathcal{P}} \binom{|p_L| + |p_R|}{|p_L|} \frac{z^{|p_L|+|p_R|+1}}{(|p_L| + |p_R| + 1)!} (ipl(p_L) + ipl(p_R) + |p_L| + |p_R|)$

Differentiate. $C'(z) = \sum_{p_L \in \mathcal{P}} \sum_{p_R \in \mathcal{P}} \frac{z^{|p_L|}}{|p_L|!} \frac{z^{|p_R|}}{|p_R|!} (ipl(p_L) + ipl(p_R) + |p_L| + |p_R|)$

Tricky; often works with perms

$$= 2C(z)P(z) + 2zP'(z)P(z) = \frac{2C(z)}{1-z} + \frac{2z}{(1-z)^3}$$

$$P(z) = \sum_{p \in P} \frac{z^{|p|}}{|p|!} = \frac{1}{1-z}$$

$$P'(z) = \sum_{p \in P} \frac{z^{|p|-1}}{(|p|-1)!} = \frac{1}{(1-z)^2}$$

CGF functional equation for path length in BSTs

$$C'(z) = \frac{2C(z)}{1-z} + \frac{2z}{(1-z)^3}$$

Look familiar?

Solving the Quicksort recurrence with OGFs

$$C_N = N + 1 + \frac{2}{N}\sum_{1\le k\le N} C_{k-1}$$

Multiply both sides by N.

$$NC_N = N(N+1) + 2\sum_{1\le k\le N} C_{k-1}$$

Multiply by z^N and sum.

$$\sum_{N\ge 1} NC_N z^N = \sum_{N\ge 1} N(N+1)z^N + 2\sum_{N\ge 1}\sum_{1\le k\le N} C_{k-1}z^N$$

Evaluate sums to get an ordinary differential equation

$$C'(z) = \frac{2}{(1-z)^3} + 2\frac{C(z)}{1-z}$$

homogeneous equation

$$\rho'(z) = 2\rho(z)/(1-z)$$

solution (integration factor)

$$\rho(z) = 1/(1-z)^2$$

Solve the ODE.

$$((1-z)^2C(z))' = (1-z)^2C'(z) - 2(1-z)C(z)$$

$$= (1-z)^2\left(C'(z) - 2\frac{C(z)}{1-z}\right) = \frac{2}{1-z}$$

Integrate.

$$C(z) = \frac{2}{(1-z)^2}\ln\frac{1}{1-z}$$

Expand.

$$C_N = [z^N]\frac{2}{(1-z)^2}\ln\frac{1}{1-z} = 2(N+1)(H_{N+1}-1)$$

21

Expected path length in BST built from a random permutation: full derivation

CGF.

$$C(z) = \sum_{p \in P} \mathrm{ipl}(p) \frac{z^{|p|}}{|p|!}$$

Decompose.

$$C(z) = \sum_{p_L \in \mathcal{P}} \sum_{p_R \in \mathcal{P}} \binom{|p_L| + |p_R|}{|p_L|} \frac{z^{|p_L|+|p_R|+1}}{(|p_L| + |p_R| + 1)!} (ipl(p_L) + ipl(p_R) + |p_L| + |p_R|)$$

Differentiate.

$$C'(z) = \sum_{p_L \in \mathcal{P}} \sum_{p_R \in \mathcal{P}} \frac{z^{|p_L|}}{|p_L|!} \frac{z^{|p_R|}}{|p_R|!} (ipl(p_L) + ipl(p_R) + |p_L| + |p_R|)$$

$$P(z) = \sum_{p \in P} \frac{z^{|p|}}{|p|!} = \frac{1}{1 - z}$$

$$P'(z) = \sum_{p \in P} \frac{z^{|p|-1}}{(|p| - 1)!} = \frac{1}{(1 - z)^2}$$

Simplify.

$$= 2C(z)P(z) + 2zP'(z)P(z)$$

$$= \frac{2C(z)}{1 - z} + \frac{2z}{(1 - z)^3}$$

Solve the ODE (see GF lecture).

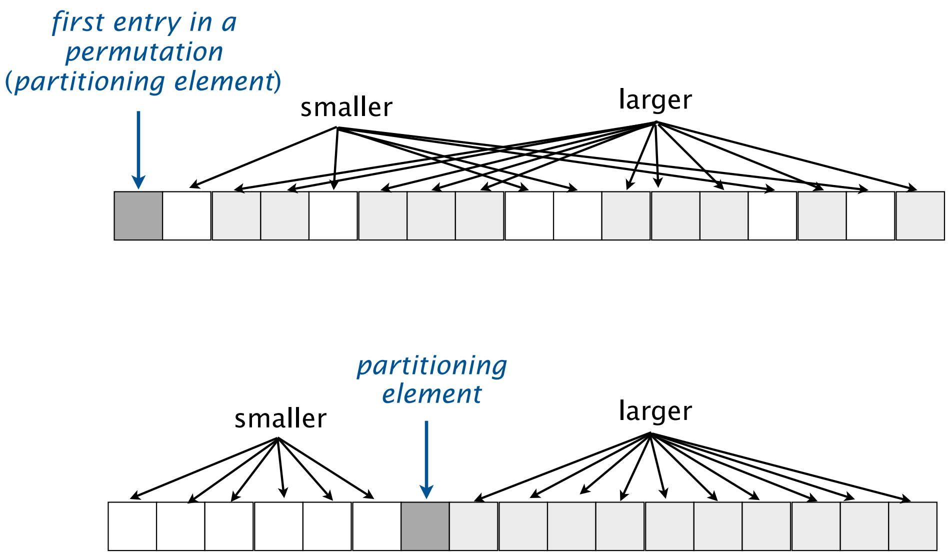
$$C(z) = \frac{2}{(1 - z)^2} \ln \frac{1}{1 - z} - \frac{2z}{(1 - z)^2}$$

Expand.

$$C_N = 2(N + 1)(H_{N+1} - 1) - 2N \sim 2N \ln N$$

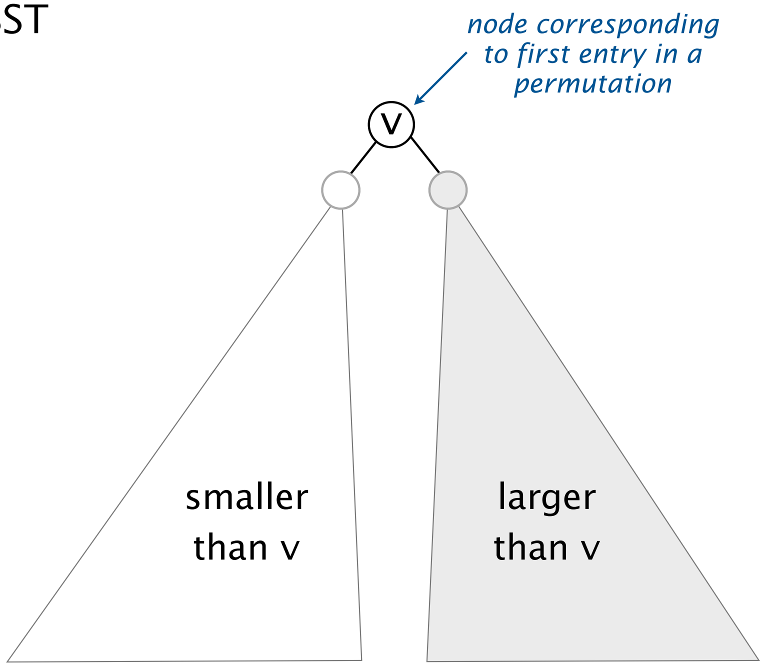
BST – quicksort bijection

Quicksort



model: random permutation
 # compares: $N+1$ + # compares for subfiles

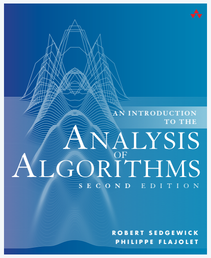
BST



model: random permutation
xpl: $N+1$ + xpl of subtrees

Average # compares for quicksort
= average external path length of BST *built from a random permutation*
= average internal path length + $2N$

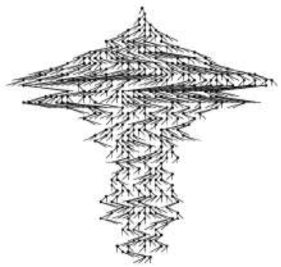
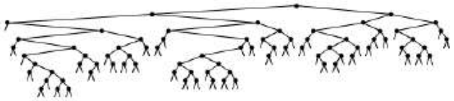
Height and other parameters



Approach works for any "additive parameter" (see text).

Height requires a different (much more intricate) approach (see text).

Summary:

	typical shape	*average path length*	*height*
random binary tree		$\sim \sqrt{\pi N}$	$\sim 2\sqrt{\pi N}$
BST built from random permutation		$\sim 2 \ln N$	$\sim c \ln N$

$c \doteq 4.311$