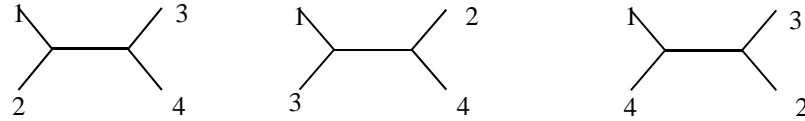


Combinatorics of Phylogenies

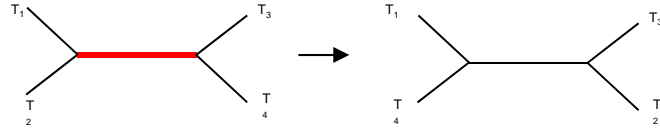
- *A list of problems*
- *Motivation*
 - *Evaluating the Size of Problem*
 - *Understanding the Structure of Problem*
 - *Designing Combinatorial Search Algorithms*
- *Topics*
 - *Enumerating main classes of trees*
 - *Enumerating other Genealogical Structures*

What is of interest to calculate ?

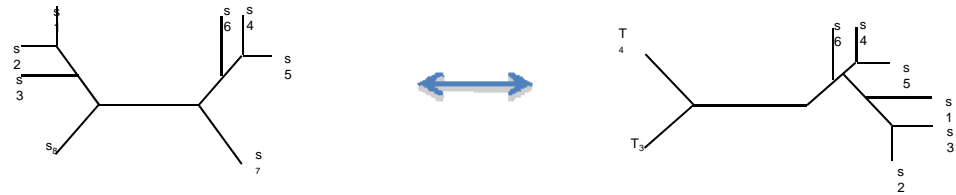
- The number of trees



- Operations on Trees



- Metrics on Trees



- Alignment of Trees

- Averages/Consensus of Trees

- Counting other genealogical structures

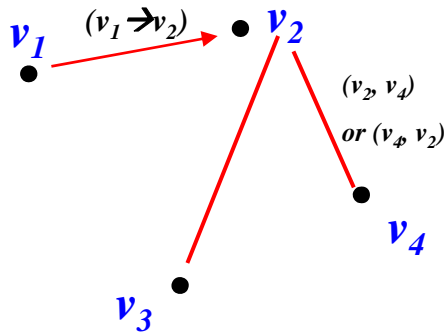
- Trees and Supertrees



- Phylogenetic Coverage

Trees – graphical & biological.

A **graph** is a set **vertices** (nodes) $\{v_1, \dots, v_k\}$ and a set of **edges** $\{e_1=(v_{i1}, v_{j1}), \dots, e_n=(v_{in}, v_{jn})\}$. Edges can be directed, then (v_i, v_j) is viewed as different (opposite direction) from (v_j, v_i) - or undirected.



Nodes can be **labelled** or **unlabelled**. In phylogenies the **leaves** are labelled and the rest unlabelled

The **degree** of a node is the number of edges it is a part of. A **leaf** has degree 1.

A graph is **connected**, if any two nodes has a path connecting them.

A **tree** is a connected graph without any cycles, i.e. only one path between any two nodes.

Trees & phylogenies.

A tree with k nodes has $k-1$ edges. (easy to show by induction)..

A **root** is a special node with degree 2 that is interpreted as the point furthest back in time. The leaves are interpreted as being contemporary.

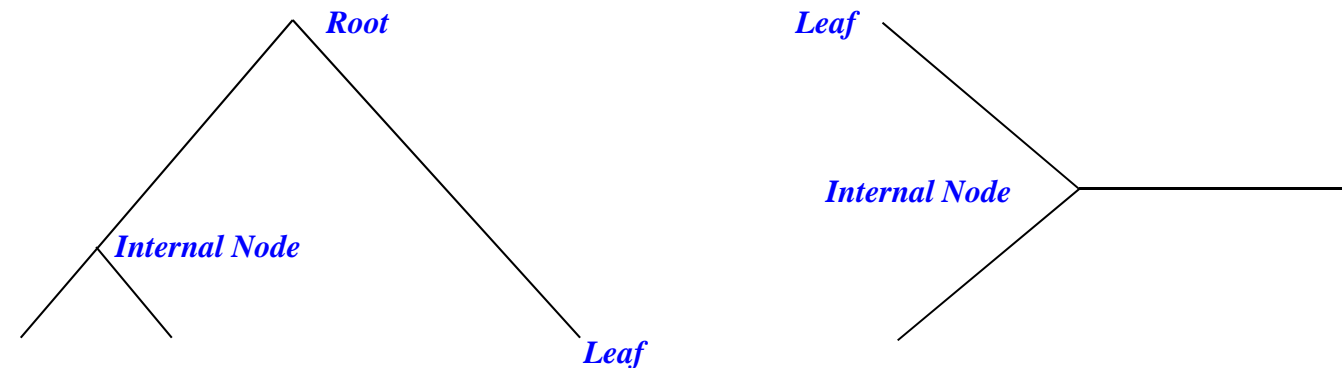
A root introduces a **time direction** in a tree.

A rooted tree is said to be **bifurcating**, if all non-leaves/roots has degree 3, corresponding to 1 **ancestor** and 2 **children**. For unrooted tree it is said to have **valency** 3.

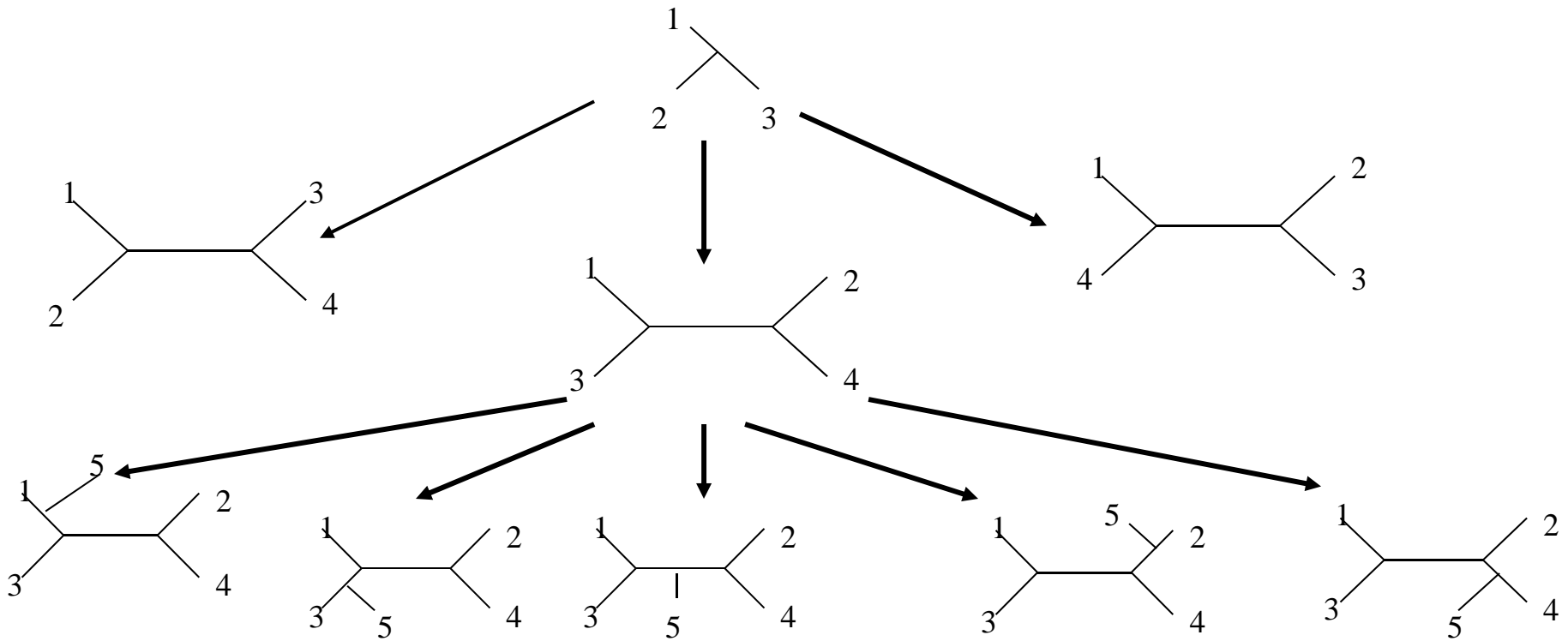
Edges can be labelled with a positive real number interpreted as **time duration** or **amount of evolution**.

If the length of the path from the root to any leaf is the same, it obeys a **molecular clock**.

Tree Topology: Discrete structure – phylogeny without branch lengths.



Enumerating Trees: Unrooted, leaflabelled & valency 3



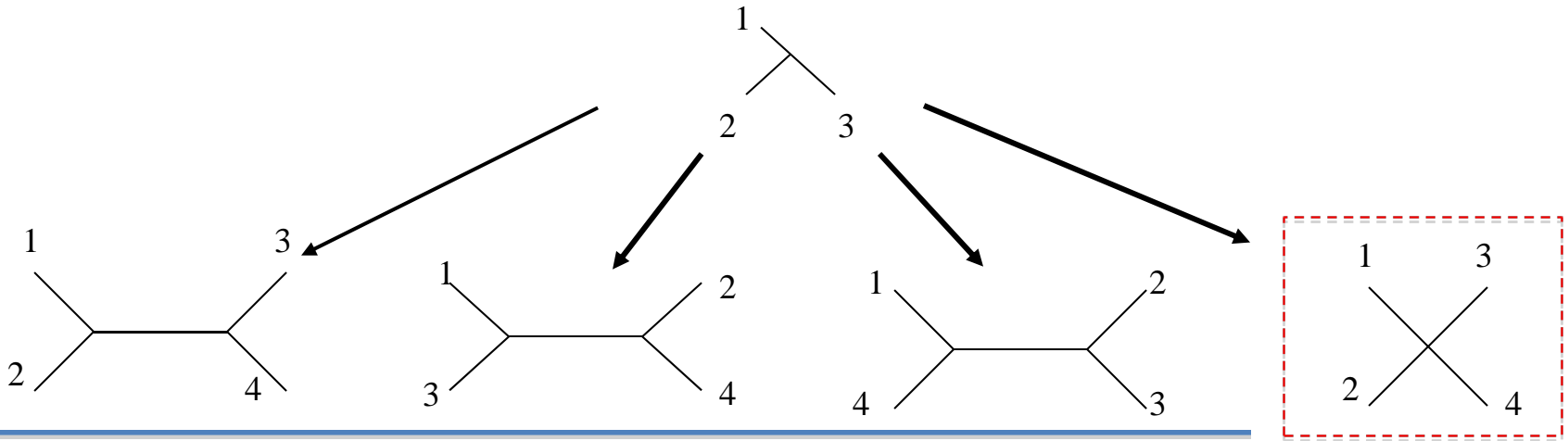
Recursion: $T_n = (2n-5) T_{n-1}$

Initialisation: $T_1 = T_2 = T_3 = 1$

$$\prod_{j=3}^{n-1} (2j-3) = \frac{(2n-5)!}{(n-3)! 2^{n-3}}$$

4	5	6	7	8	9	10	15	20	
3	15	105	945	10345	1.4 10 ⁵	2.0 10 ⁶	7.9 10 ¹²	2.2 10 ²⁰	

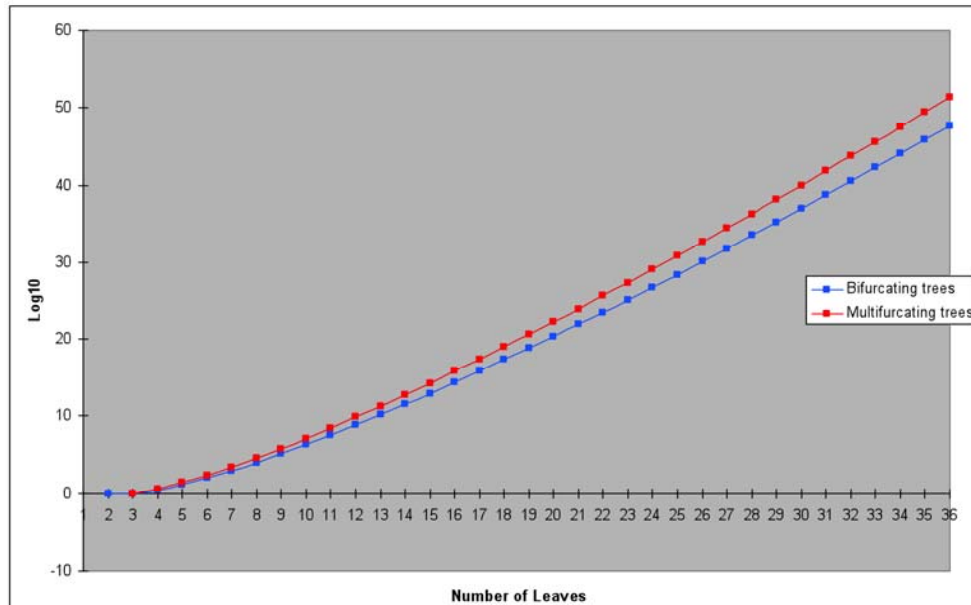
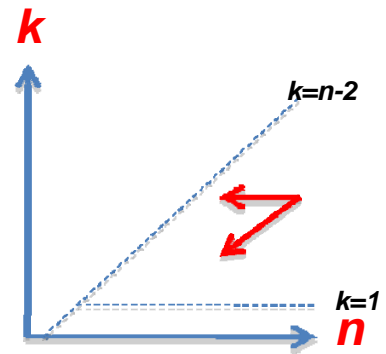
Number of leaf labelled phylogenies with arbitrary valencies



• n – number of leaves, k – number of internal nodes

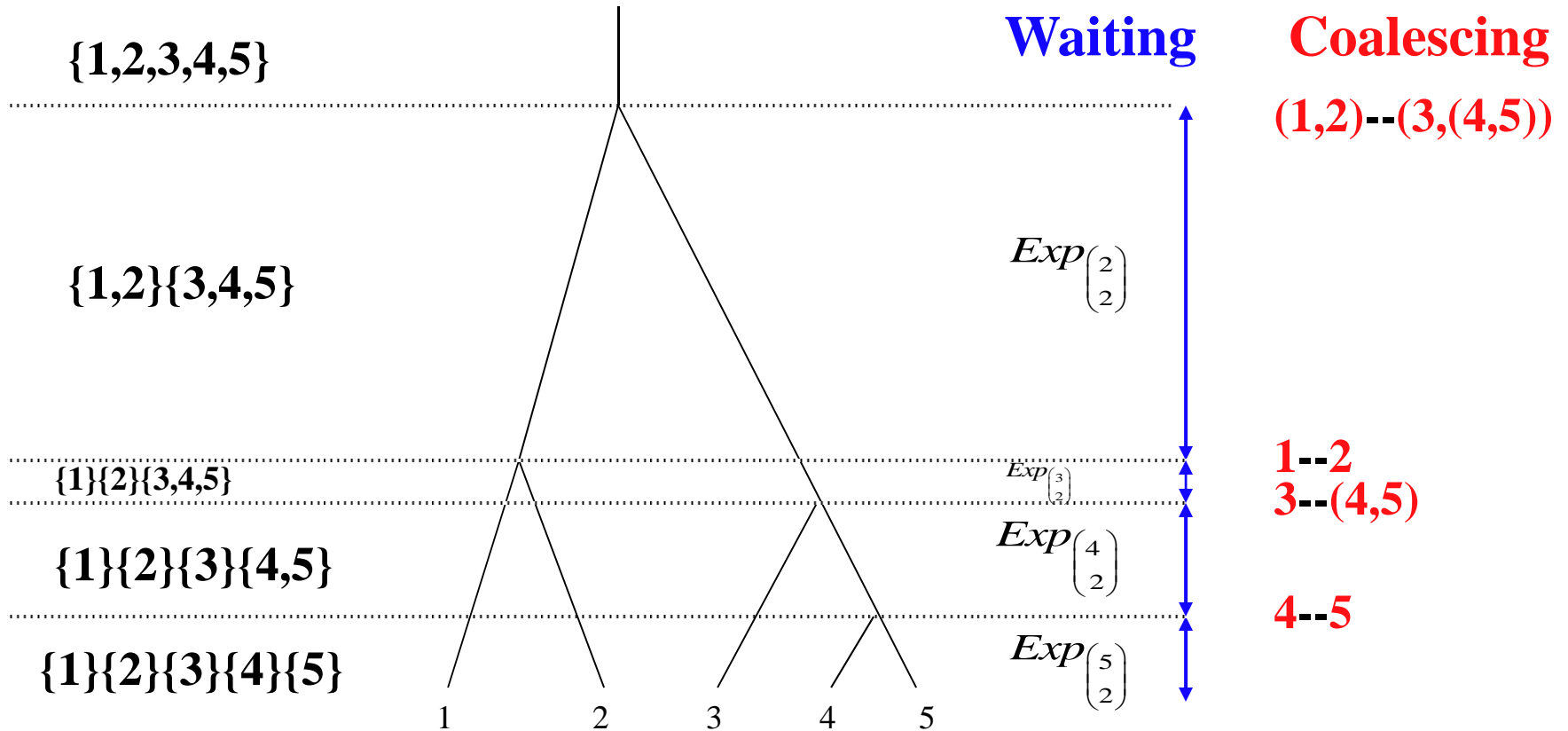
Recursion: $R_{n,k} = (n+k-3) R_{n-1,k-1} + k R_{n-1,k}$

Initialisation: $R_{n,1} = 1, R_{n,n-2} = T_n$



Number of Coalescent Topologies

- Time ranking of internal nodes are recorded



- Bifurcating:**

$$S_1 = S_2 = 1$$

$$S_j = \binom{j}{2} S_{j-1}$$

$$S_n = \prod_{j=2}^n \binom{j}{2} = \frac{j!(j-1)!}{2^{j-1}}$$

- Multifurcating:**

$$Q_j = \sum_{i=1}^{j-1} \text{Stirling}[j, i] Q_i$$

Unlabelled counting: Sketch of method

Let g_n be the size of a class index by n – for instance number of trees with n nodes. The function

$G(z) = \sum_{i=0}^{\infty} g_n z^n$ is called the generating function and is central in counting trees and much more

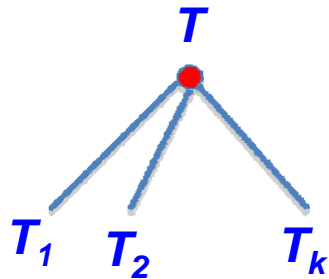
For certain recursive structures, the counting problem can be rephrased as functional equations in G

If any combinatorial object a_k from A_n , can be written as (b_i, c_j) [b_i from B and c_j from C]. Then

$G_A = G_B * G_C$, since $a_k = b_1 c_{k-1} + \dots + b_{k-1} c_1$

$k-1$	$b_1 c_{k-1}$		
2	$b_1 c_2$	$b_2 c_2$	
1	$b_1 c_1$	$b_2 c_1$	$b_{k-1} c_1$
	1	2	$k-1$

Rooted trees, ordered subtrees of arbitrary degree:



$$G(z) = z \sum_{i=0}^{\infty} G(z)^i = z / (1 - G(z))$$

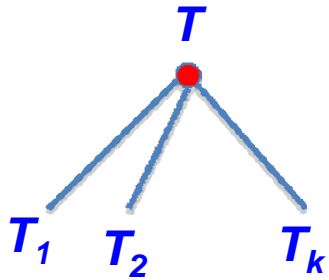
$$G(z) = [1 - \sqrt{1 - 4z}] / 2$$

$$g_n = 1/n \binom{2n-2}{n-1}$$

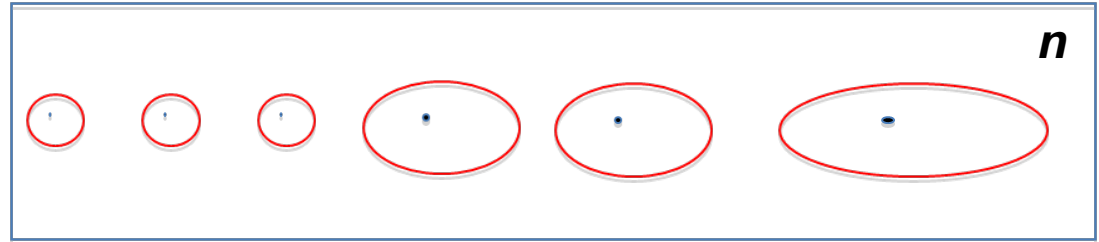
Equivalent to set of nested parenthesis, whose size is described by the Catalan numbers

Sketch of the problems: Multifurcations

rooted trees, unordered subtrees



Since tree class can occur in multiplicities, counting must be done accordingly corresponding to the simple case in the bifurcating case where left and right subtree had the same size.



How many ways can n be partitioned? The above: $i^3j^2k^1$ $[3i+2j+k=n]$

How many integers occur in multiplicities?

Within a multiplicity, how many ways can you choose unordered tuples?

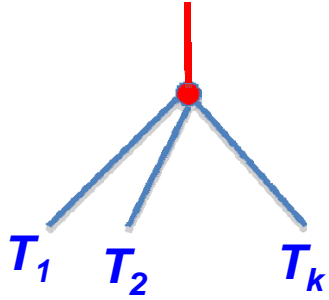


Functional Equation: $G(z) = z \text{Exp}(G(z))$

Asymptotically: $\lambda \beta^n n^{-3/2}$ $\lambda \sim .43992$ $\beta \sim 2.95576$

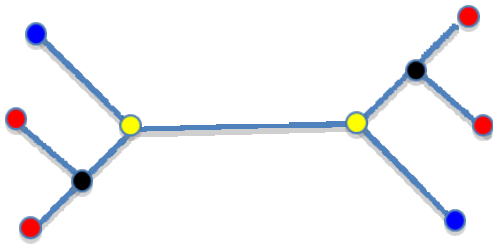
Sketch of the problems: De-rooting

De-rooting

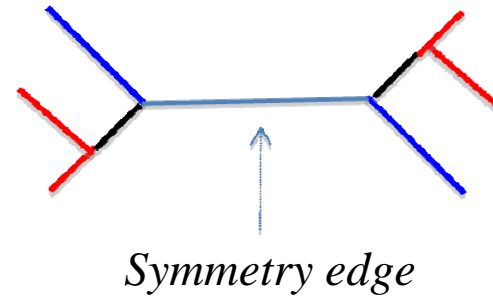


If the root is removed, trees that are different when the root is known, can become identical.

Set of dissimilar nodes (4)



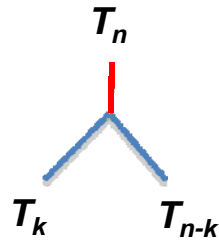
Set of dissimilar edges (3)



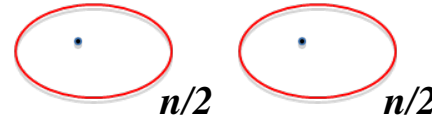
Node classes – edges classes [ignoring symmetry edge] = 1 for any unlabelled, unrooted tree

Counting rooted unordered bifurcating trees

Recursive argument:



Each choice of Q from and P from will create new T except when $k=n-k$.



$$T_n = \sum_1^{[(n+1)/2]} T_k T_{n-k} \text{ if } n \text{ odd} \quad T_n = \sum_1^{[(n-1)/2]} T_k T_{n-k} + T_{n/2}(T_{n/2} + 1)/2 \text{ if } n \text{ even} \quad T_1 = 1$$

1 1

2 : — 1

3 • — • 1

4 • — • — • 2

5 • — • — • — • 3

6 • — • — • — • — • 6

7 • — • — • — • — • — • 11

Functional Equation:

$$G(z) = z + [G(z)^2 + G(z^2)]/2$$

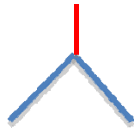
Asymptotically:

$$\lambda \beta^n n^{-3/2} \quad \lambda \sim .31877 \quad \beta \sim 2.48325$$

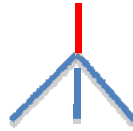
Numbers of tree shapes

from Felsenstein, 2003

Number
of Leaves



Rooted bifurcating



Rooted multifurcating



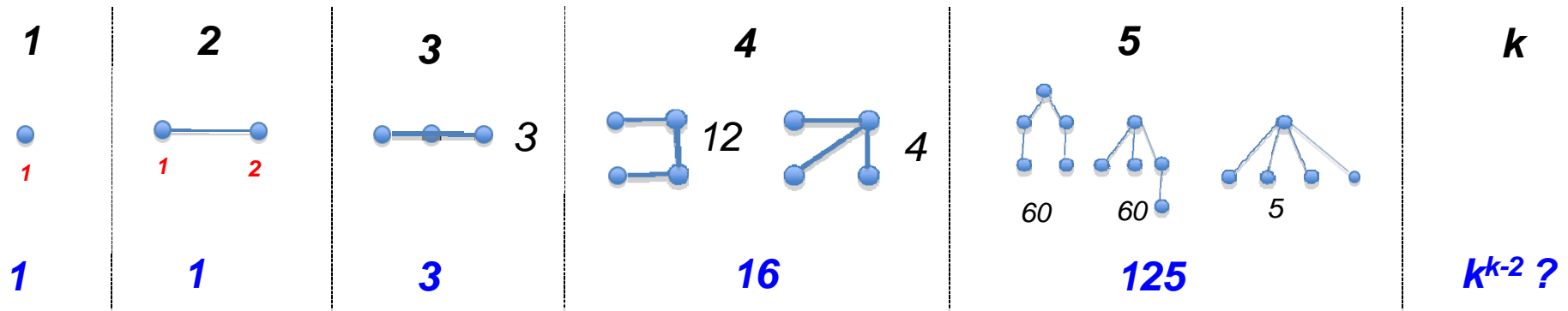
Unrooted bifurcating



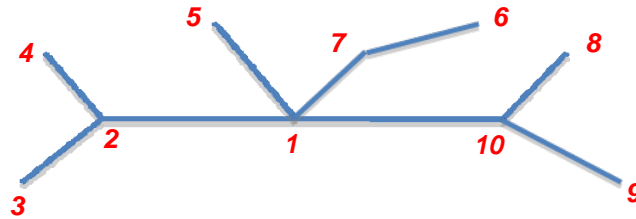
Unrooted multifurcating

1	1	1	1	1
2	1	1	1	1
3	1	2	1	1
4	2	5	1	2
5	3	12	1	3
6	6	33	2	7
7	11	90	2	13
8	23	261	4	33
9	46	766	6	73
10	98	2.312	12	202
11	207	7.068	18	488
12	451	21.965	41	1.441
13	983	68.954	66	3.741
14	2.179	218.751	154	11.496
15	4.850	699.534	265	31.311
16	10.905	2.253.676	628	98.607
17	24.631	7.305.676	1.132	278.840
18	56.011	23.816.743	2.748	895.137
19	127.912	78.023.602	5.098	2.599.071
20	293.547	256.738.751	12.444	8.452.620

Pruefer Code: Number of Spanning trees on labeled nodes



Proof by Bijection to $k-2$ tuples of $[1, \dots, k]$ (Pruefer 1918): From van Lint and Wilson



From tree to tuple:

Remove leaf with lowest index	b_i	3	4	2	5	6	7	1	8
Register attachment of leaf	a_i	2	2	1	1	7	1	10	10

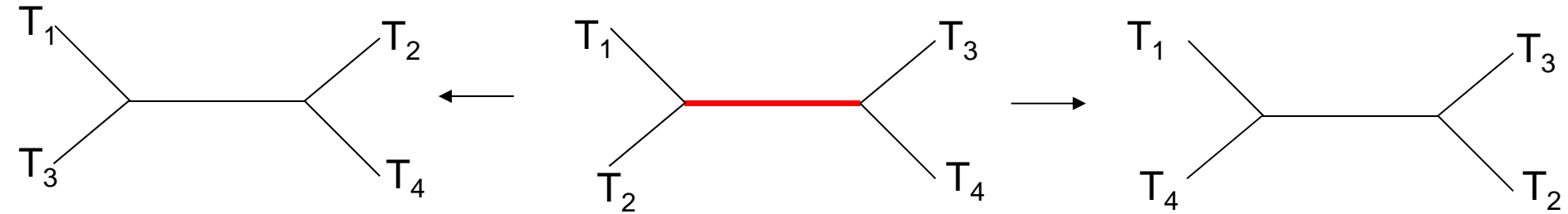
From tuple to tree: Given a_1, \dots, a_{n-2} , set $a_{n-1} = n$

Let b_i be smallest $\{a_i, a_{i+1}, \dots, a_{n-1}\} \cup \{b_1, b_2, \dots, b_{i-1}\}$

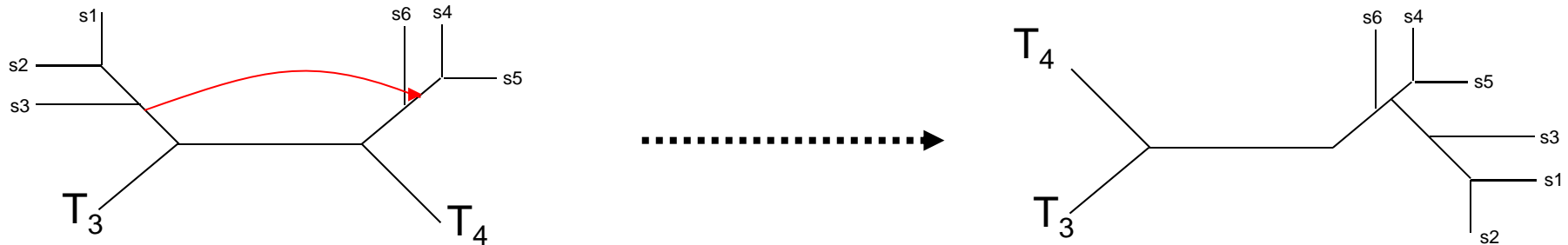
Then $\{(b_i, a_i) : i=1, \dots, n-1\}$ will be the edge set of the spanning tree

Heuristic Searches in Tree Space

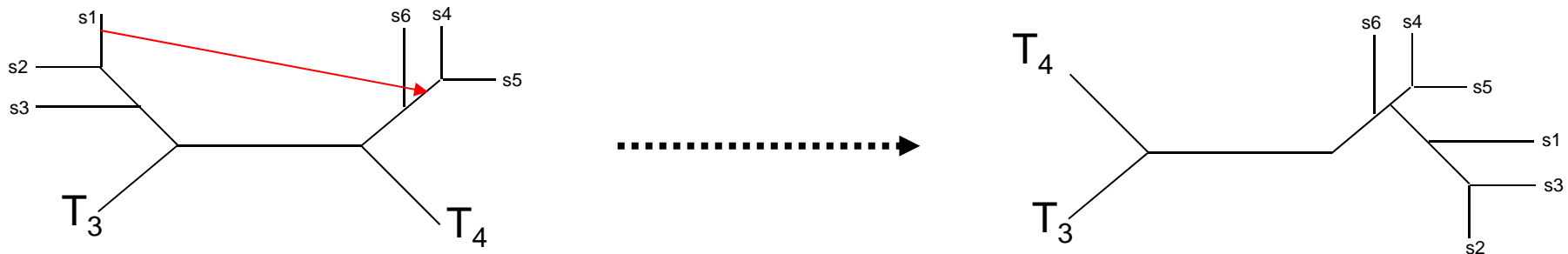
Nearest Neighbour Interchange



Subtree regrafting



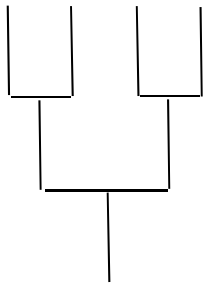
Subtree rerooting and regrafting



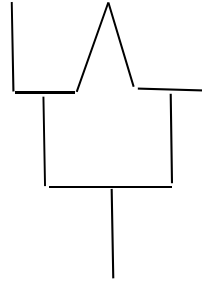
Counting Pedigrees

1 extant individual, discrete generations, ancestors sex-labelled?:

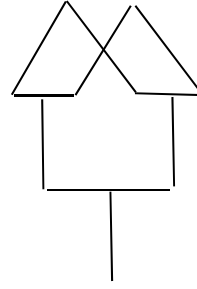
2
1
0



1



2



1

3

4

Counting Sex-Labelled Pedigrees

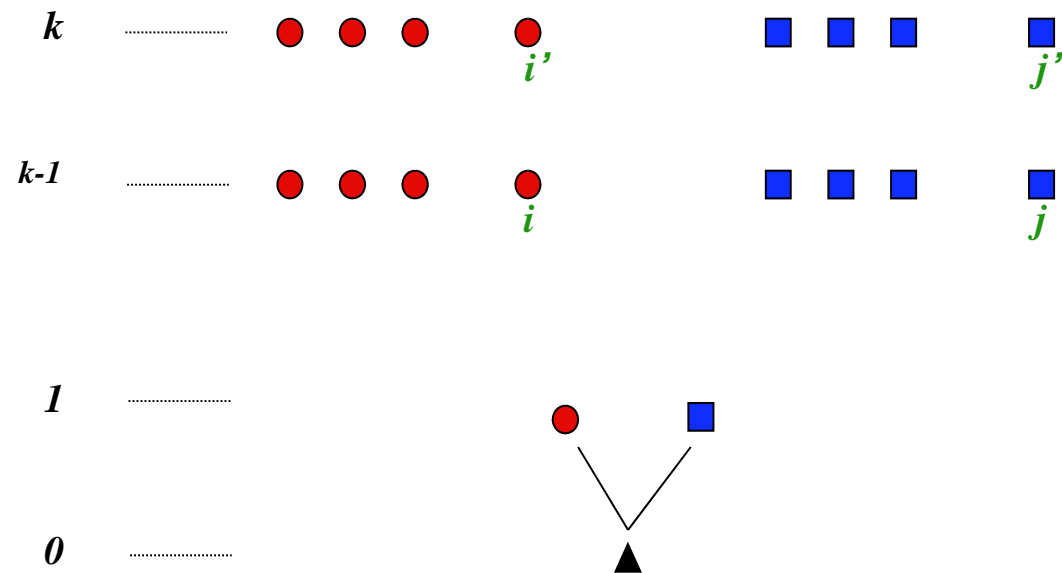
Tong Chen & Rune Lyngsø

$A_k(i,j)$ - the number of pedigrees k generations back with i females, k males

$S(n,m)$ - Stirling numbers of second kind - ways to partition n labeled objects into m unlabelled groups.

Recursion:

$$A_k(i', j') = \sum A_{k-1}(i, j) S_{k-1}(i + j, i') S_{k-1}(i + j, j')$$



2	4
3	279
4	$2.8 \cdot 10^7$
5	$2.8 \cdot 10^{20}$
6	$7.4 \cdot 10^{52}$
7	$2.8 \cdot 10^{131}$
8	$2.9 \cdot 10^{317}$
9	$3.5 \cdot 10^{749}$
10	$3.9 \cdot 10^{1737}$

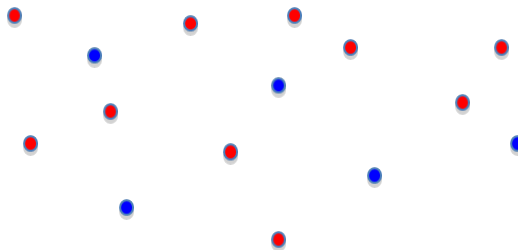
Combinatorics of Genealogical Structures

- *A list of problems*
- *A little about graphs*
- *Enumerating main classes of trees*
 - *Unrooted trees with leaves labelled*
 - *Coalescent topologies*
 - *Tree shapes*
- *Operations on Trees*
- *Enumerating other Genealogical Structures*
 - *Pedigrees*

Combinatorics of Trees and Molecules

- **1875 Cayley's enumeration of Alkanes (with brutto formula C_nH_{2n+2})**
Rains and Sloane (1999) "On Cayley Enumeration of Alkanes" J. of Integer Sequences 2
- **1937 Polya enumeration of Molecules**
G. Pólya; R. C. Read (1987). *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds*. New York: [Springer-Verlag](#). [MR0884155](#). [ISBN 0-387-96413-4](#).
- **1948 Otter asymptotic enumeration of Alkanes/Trees**
Otter "The Number of Trees" *he Annals of Mathematics, Second Series, Vol. 49, No. 3 (Jul., 1948), pp. 583-599*
- **2005 → Reymond.. exhaustive enumeration of small molecules with C, N, O, S, P.**

Sampling:



Some good looking papers/books:

Harary and Palmer (1973) *Graphical Enumeration* Acad. Press.

Sloane and Nambi (2006) "Integer Sequences Related to Chemistry"

Faulon, Visco and Roe (2005) "Enumerating Molecules" **Rev. Compu. Chem, Vol21**

Randi (2004) "Nenad Trinajstić – Pioneer of Chemical Graph Theory" *CROATICA CHEMICA ACTA* 77 (1–2) 1-15

Meringer (2010) "Structure Enumeration and Sampling" in *Handbook of Chemoinformatics Algorithms* CRC Press

Nenad Trinajstić (1992) *Chemical Graph Theory* CRC Press