

My wish for the project-examination

- *It is expected to be 3 days worth of work.*
- *You will be given this in week 8*
- *I would expect 7-10 pages*
- *You will be given 2-4 key references*
- *A set of guiding questions that might help you in your writing*
- *You can chose between a set of topics broadly covering the taught material*

"Where a topic is assessed by a mini-project, the mini-project should be designed to take a typical student about three days. You are not permitted to withdraw from being examined on a topic once you have submitted your mini-project to the Examination Schools."

- ***I emphasize – this is not formal as it has not been cleared with the appropriate committee***

Combinatorics of Phylogenies

- ***Motivation***

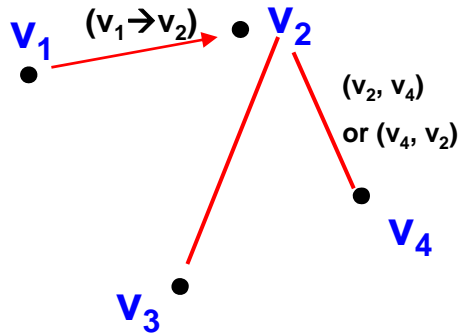
- ***Evaluating the Size of Problem***
- ***Understanding the Structure of Problem***
- ***Designing Combinatorial Search Algorithms***

- ***Topics***

- ***Enumerating main classes of trees***
- ***Enumerating other Genealogical Structures***
- ***Size of Neighborhoods***

Trees – graphical & biological.

A **graph** is a set **vertices** (nodes) $\{v_1, \dots, v_k\}$ and a set of **edges** $\{e_1=(v_{i_1}, v_{j_1}), \dots, e_n=(v_{i_n}, v_{j_n})\}$. 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.

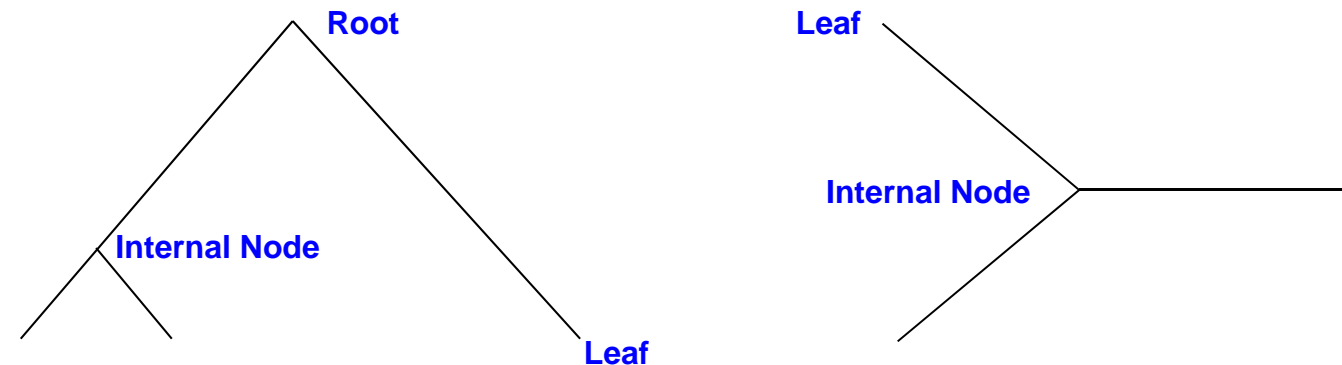
A root introduces a **time direction** in a tree.

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**.

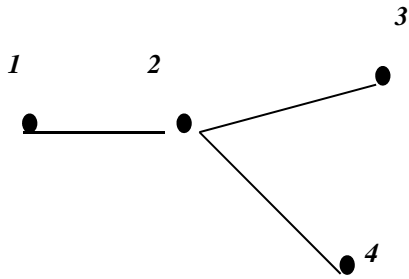
Edges can be labelled with a positive real number interpreted as **time duration** or **amount of evolution**.

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

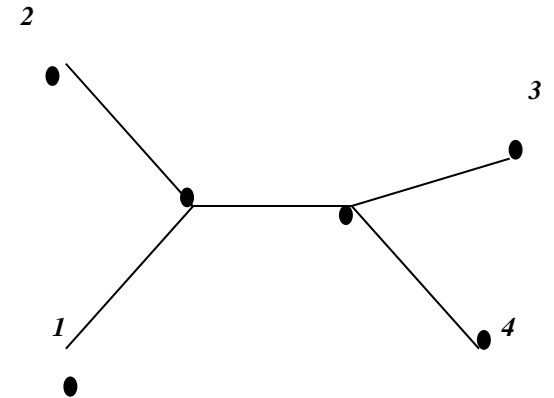
Tree Topology: Discrete structure – phylogeny without branch lengths.



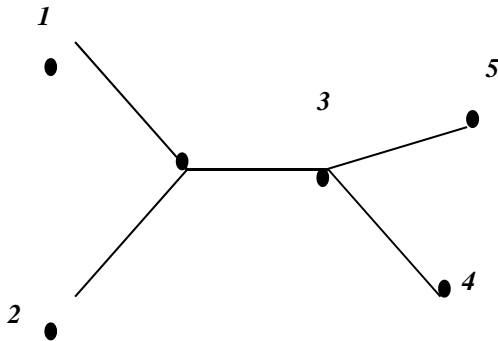
Spanning Trees, Steiner Trees & Spannoids



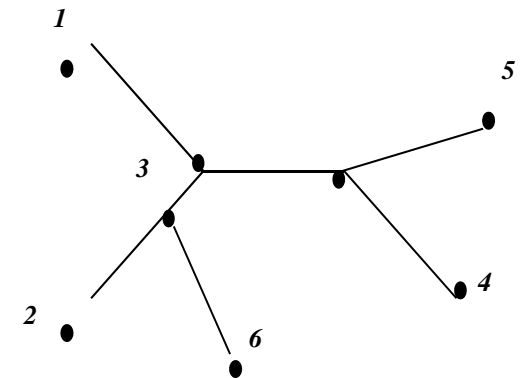
Spanning tree



Steiner tree



1-Spannoid



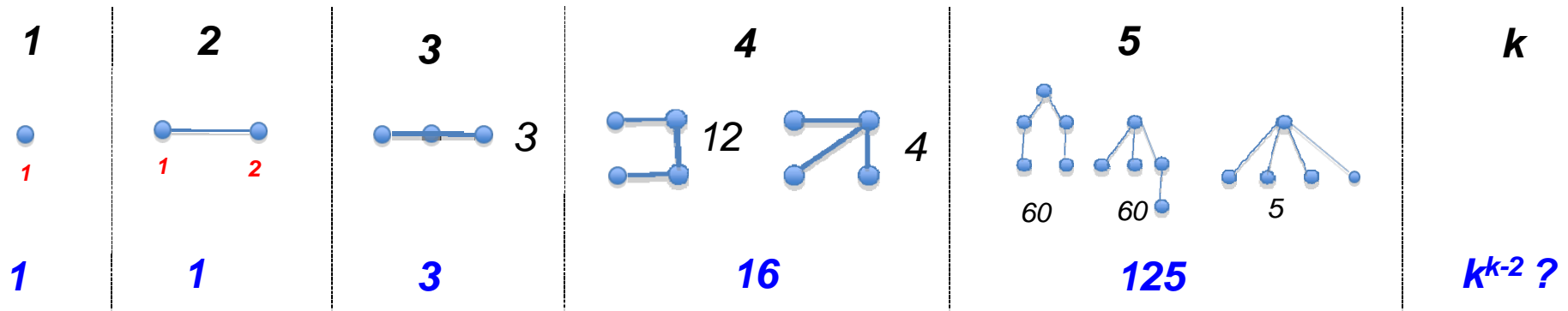
2-Spannoid

Advantage: Decomposes large trees into small trees

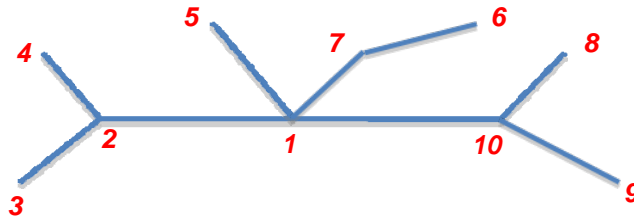
Questions: How to find optimal spannoid?

How well do they approximate?

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:

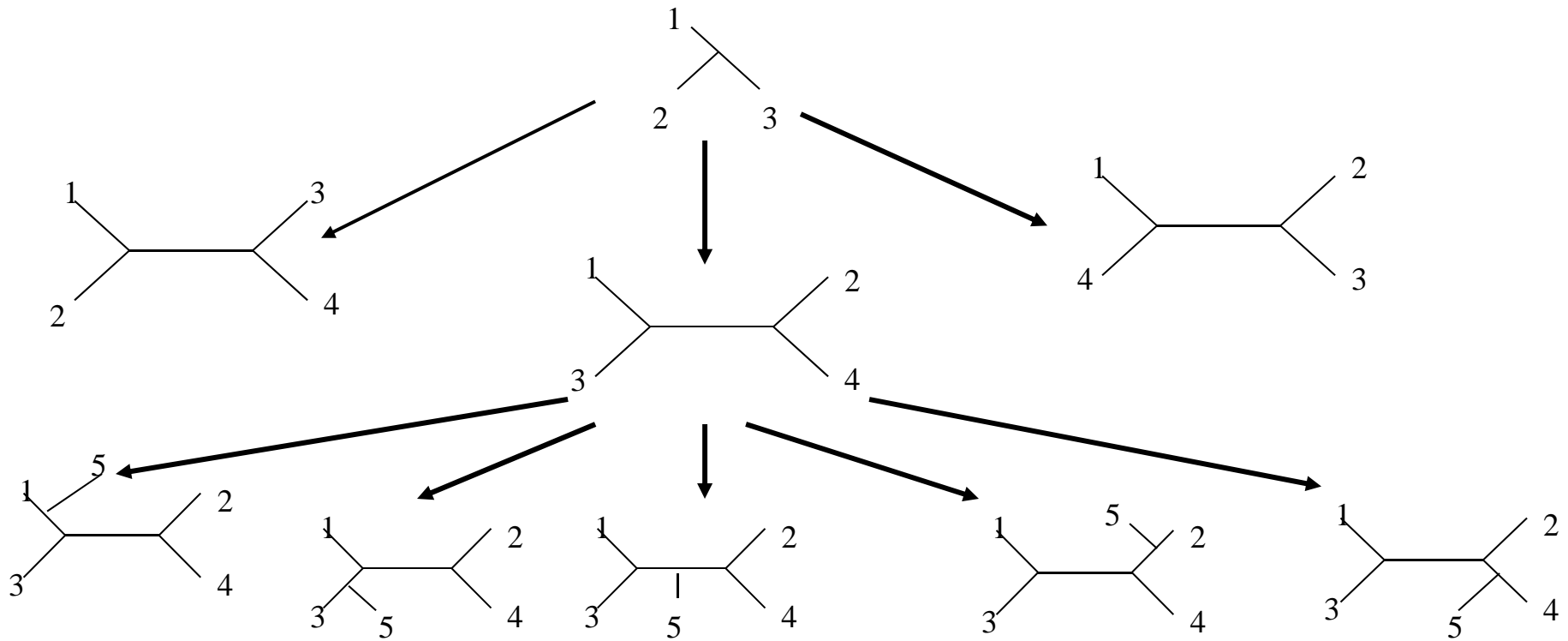
Remove leaf with lowest index	b_i	3	4	2	5	6	7	1	8
Register attachment of leaf	a_i	2	2	1	1	7	1	10	10

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

Enumerating Trees: Unrooted & valency 3



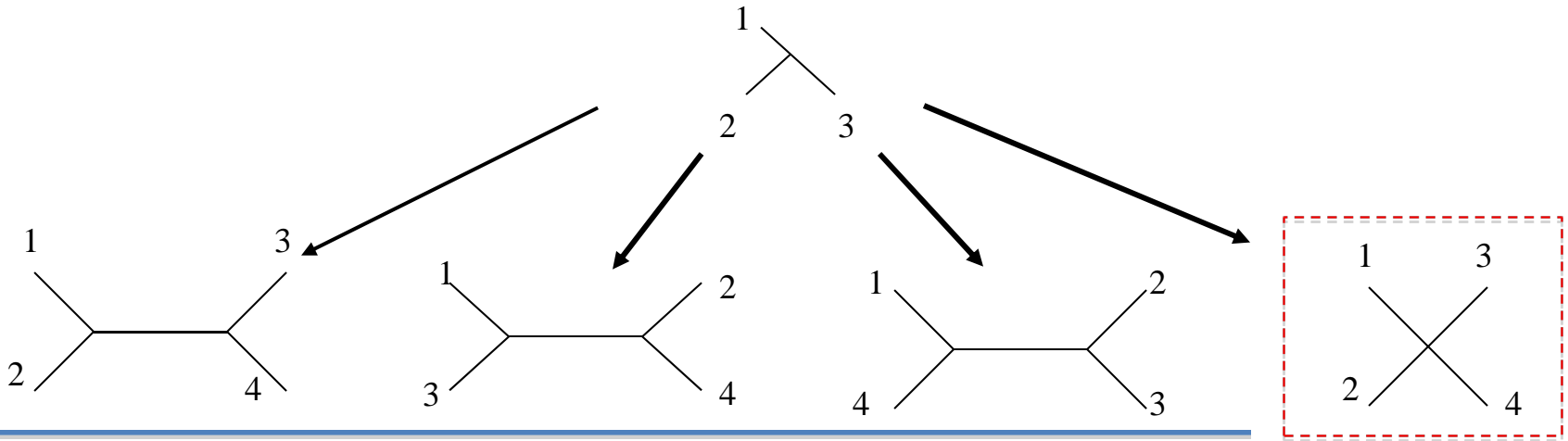
Recursion: $T_n = (2n-5) T_{n-1}$

Initialisation: $T_1 = T_2 = T_3 = 1$

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

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	

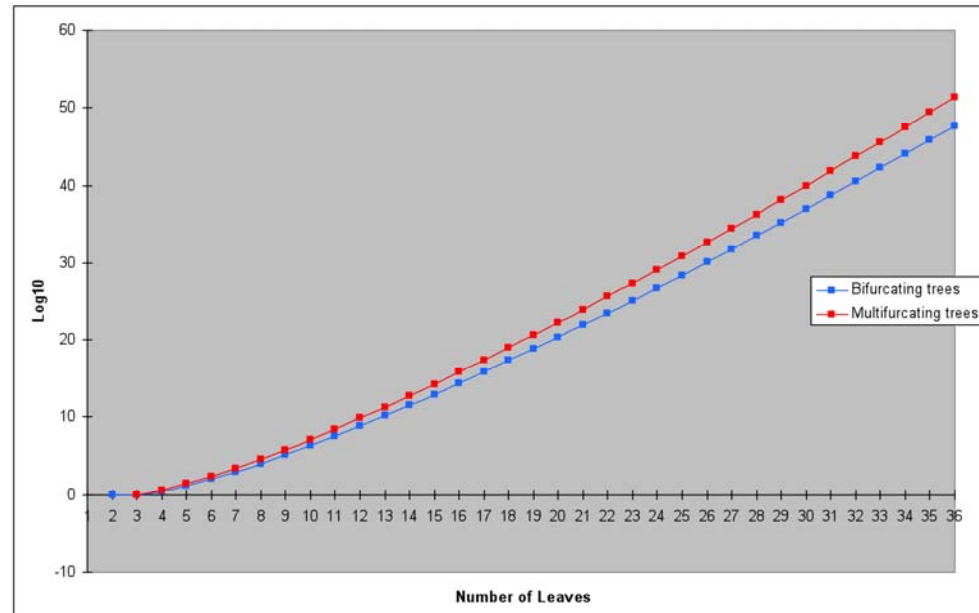
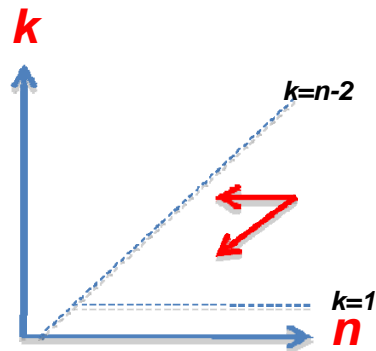
Number of phylogenies with arbitrary valencies



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

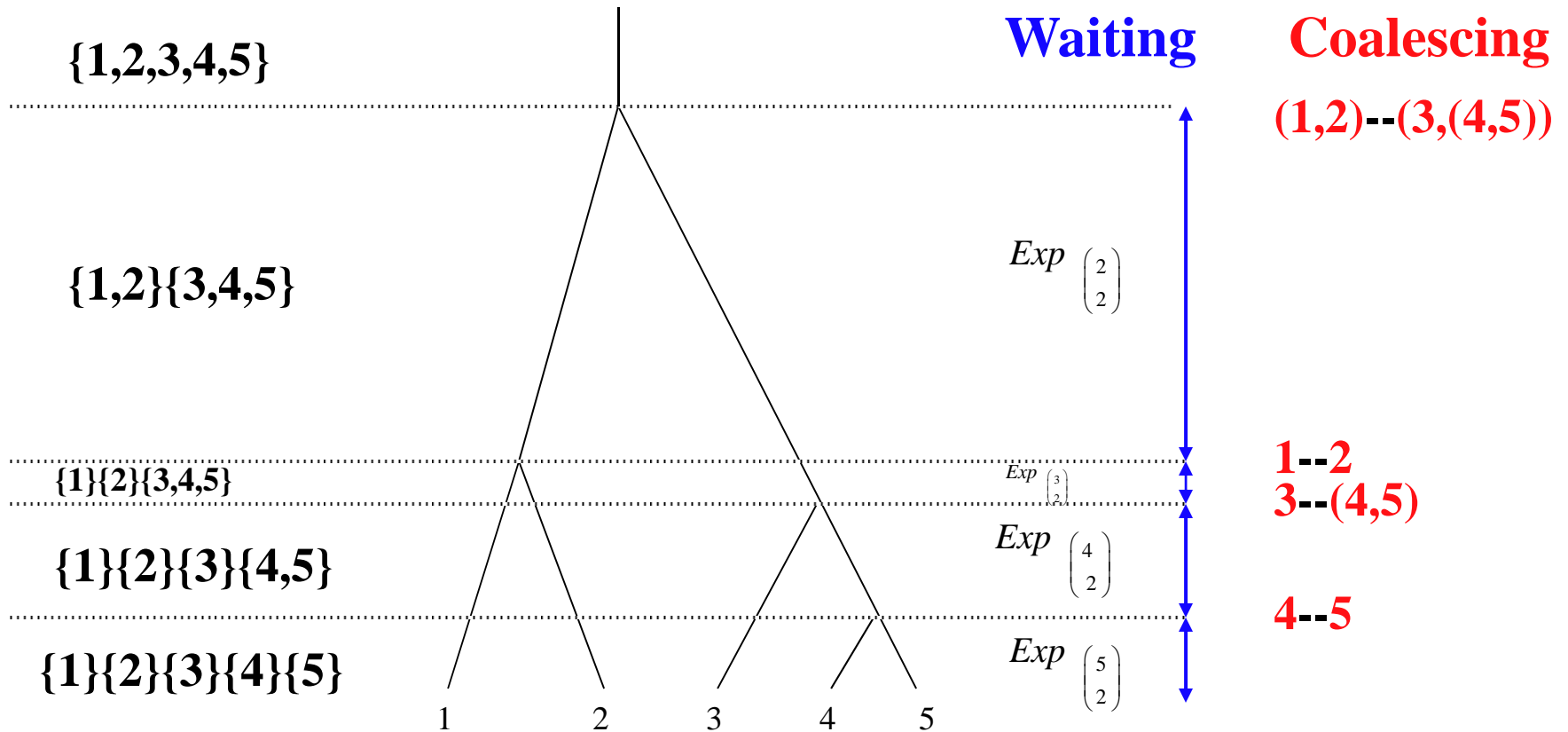
Recursion: $R_{n,k} = (n+k-3) R_{n-1,k-1} + k R_{n-1,k}$

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



Number of Coalescent Topologies

- Time ranking of internal nodes are recorded



- Bifurcating:**

$$S_1 = S_2 = 1$$

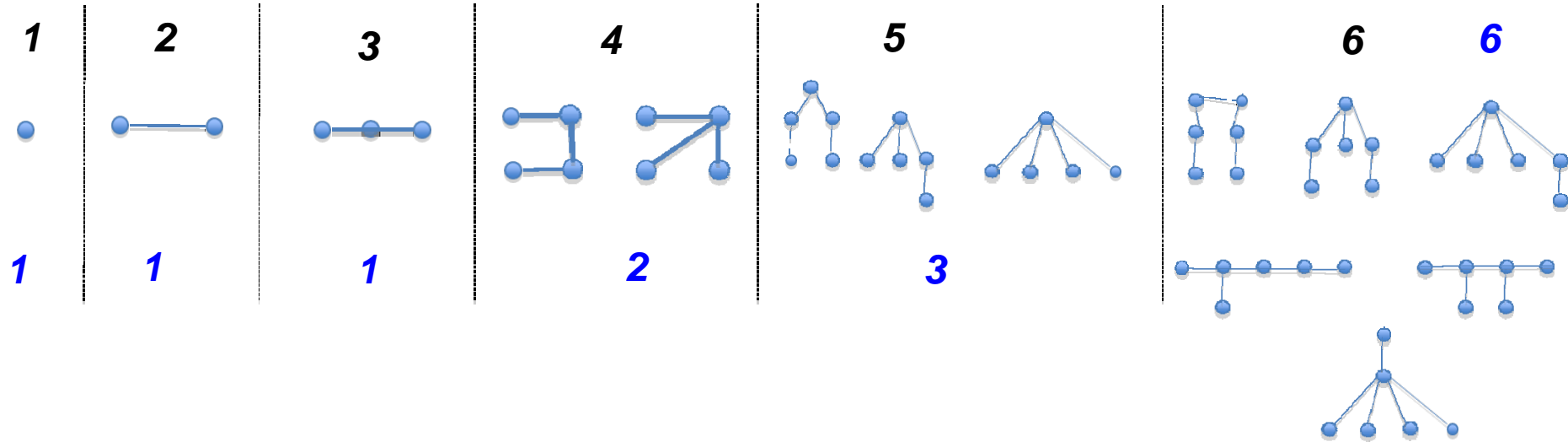
$$S_j = \binom{j}{2} S_{j-1}$$

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

- Multifurcating:**

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

Non-isomorphic trees



Counting Sex-Labelled Pedigrees

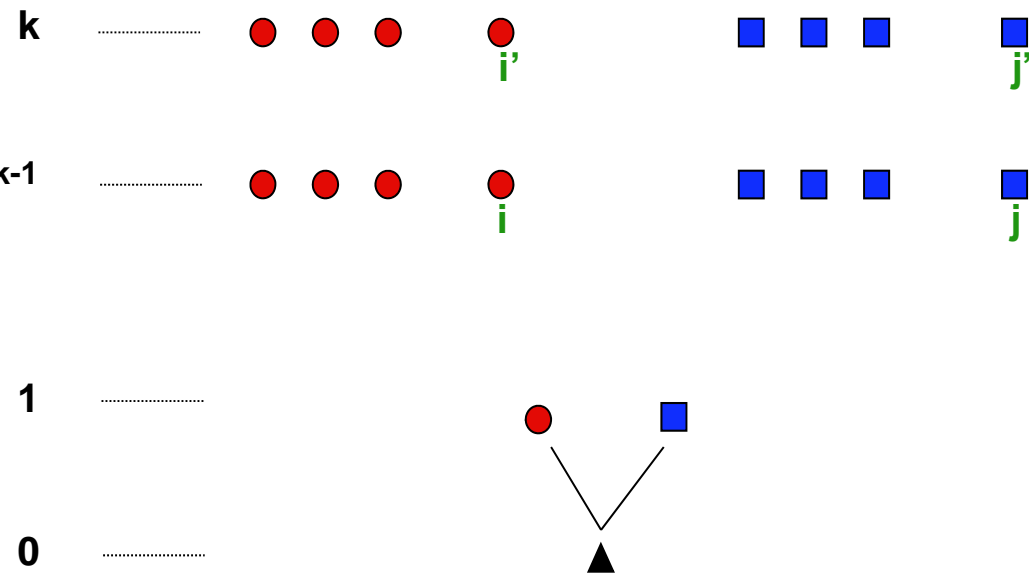
Tong Chen & Rune Lyngsø

$A_k(i,j)$ - 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:

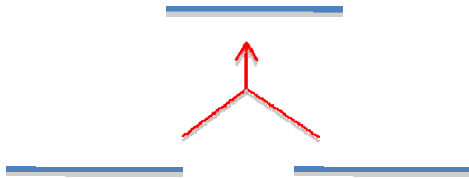
$$A_k(i',j') = \sum A_{k-1}(i,j) S_{k-1}(i+j,i') S_{k-1}(i+j,j')$$



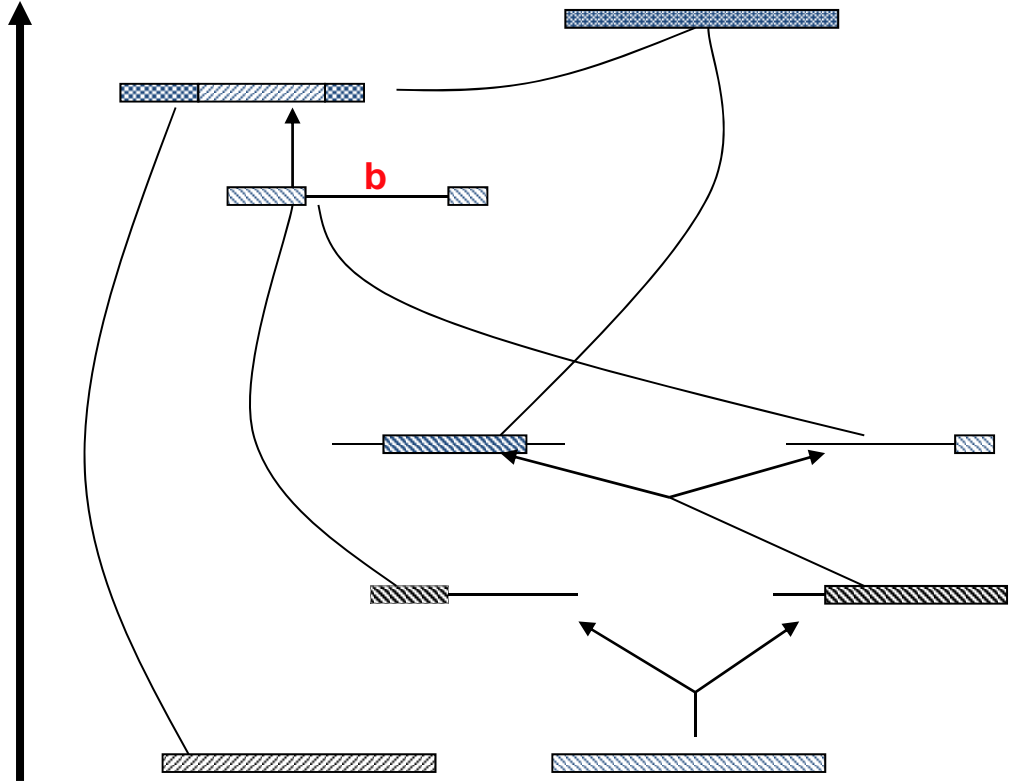
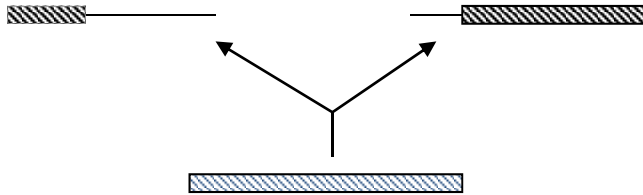
2	4
3	279
4	$2.8 * 10^7$
5	$2.8 * 10^{20}$
6	$7.4 * 10^{52}$
7	$2.8 * 10^{131}$
8	$2.9 * 10^{317}$
9	$3.5 * 10^{749}$
10	$3.9 * 10^{1737}$

Counting Ancestral Recombination Graph (ARG) Topologies

- **Coalescent/Duplication**



- **Recombination**

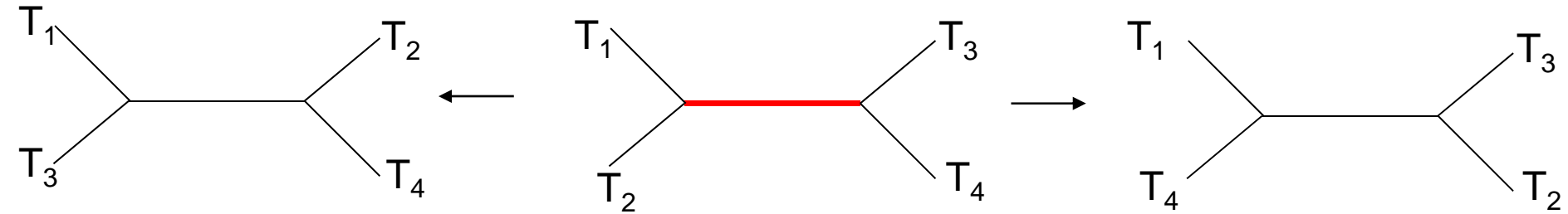


- **Each position on the sequence has a tree**
- **Neighboring positions have trees differing by at most one SPR**
- **Recombinations create time ordering**

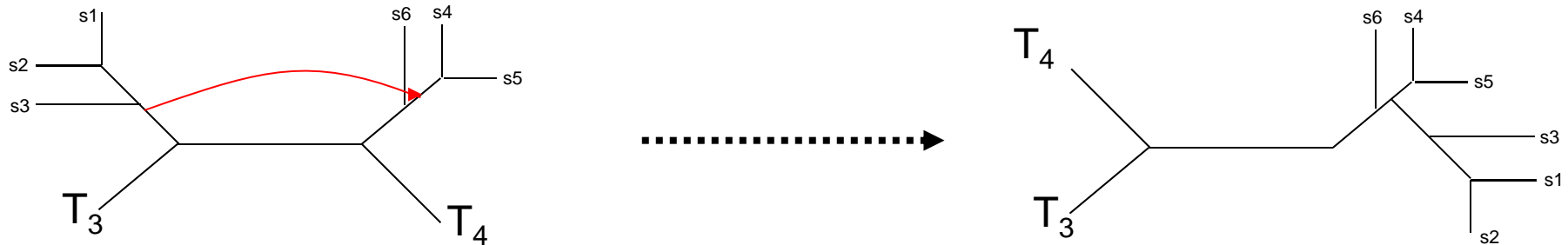
- **How is ARG topology defined**
- **How many are there?**

Heuristic Searches in Tree Space

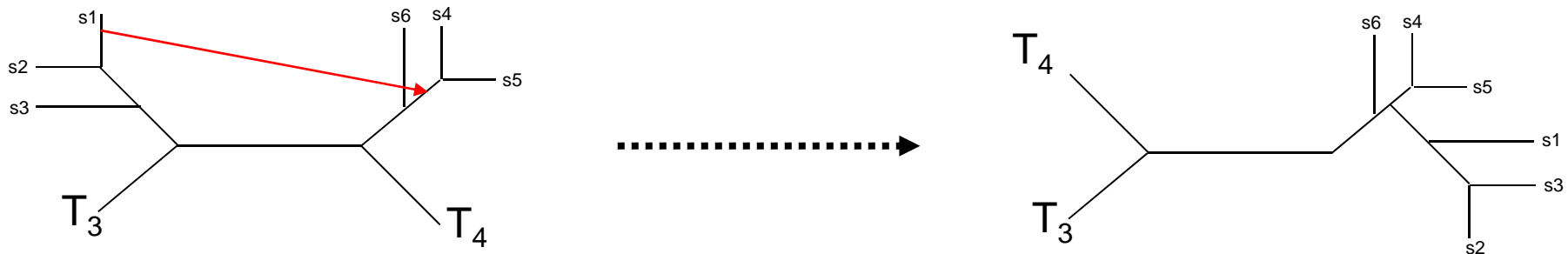
Nearest Neighbour Interchange



Subtree regrafting



Subtree rerooting and regrafting



Tree Combinatorics and Neighborhoods

Observe that the size of the unit-neighbourhood of a tree does not grow nearly as fast as the number of trees

$\delta(T)$:= number of trees one SPR operation away from a given tree T .

	Unrooted		Rooted			Dendrograms		
n	# of trees	δ	# of trees	δ_{\max}	δ_{\min}	# of trees	δ_{\max}	δ_{\min}
4	3	2	15	12	10	18	12	13
5	15	12	105	28	24	180	33	37
6	105	30	945	52	44	2,700	71	79
7	945	56	10,395	84	70	56,700	128	143
8	10,395	90	135,135	124	102	1,587,600	210	233
9	135,135	132	2,027,025	170	140	57,153,600	?	?
10	2,027,025	182	34,459,425	224	184	2,571,912,000	?	?

Due to Yun Song

$(2n-5)!! = \frac{(2n-3)!}{2^{n-1}(n-2)!}$
 $2(n-3)(2n-7)$
 $3n^2 - 13n + 14$
 $4(n-2)^2 - 2 \sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor$
 $\frac{n!(n-1)!}{2^{n-1}}$
 $\frac{1}{3} (2n^3 - 3n^2 - 20n + 39)$

Allen & Steel (2001)

Song (2003+)