

Triplet and Quartet Distances Between Trees of Arbitrary Degree

Gerth Stølting Brodal

Aarhus University

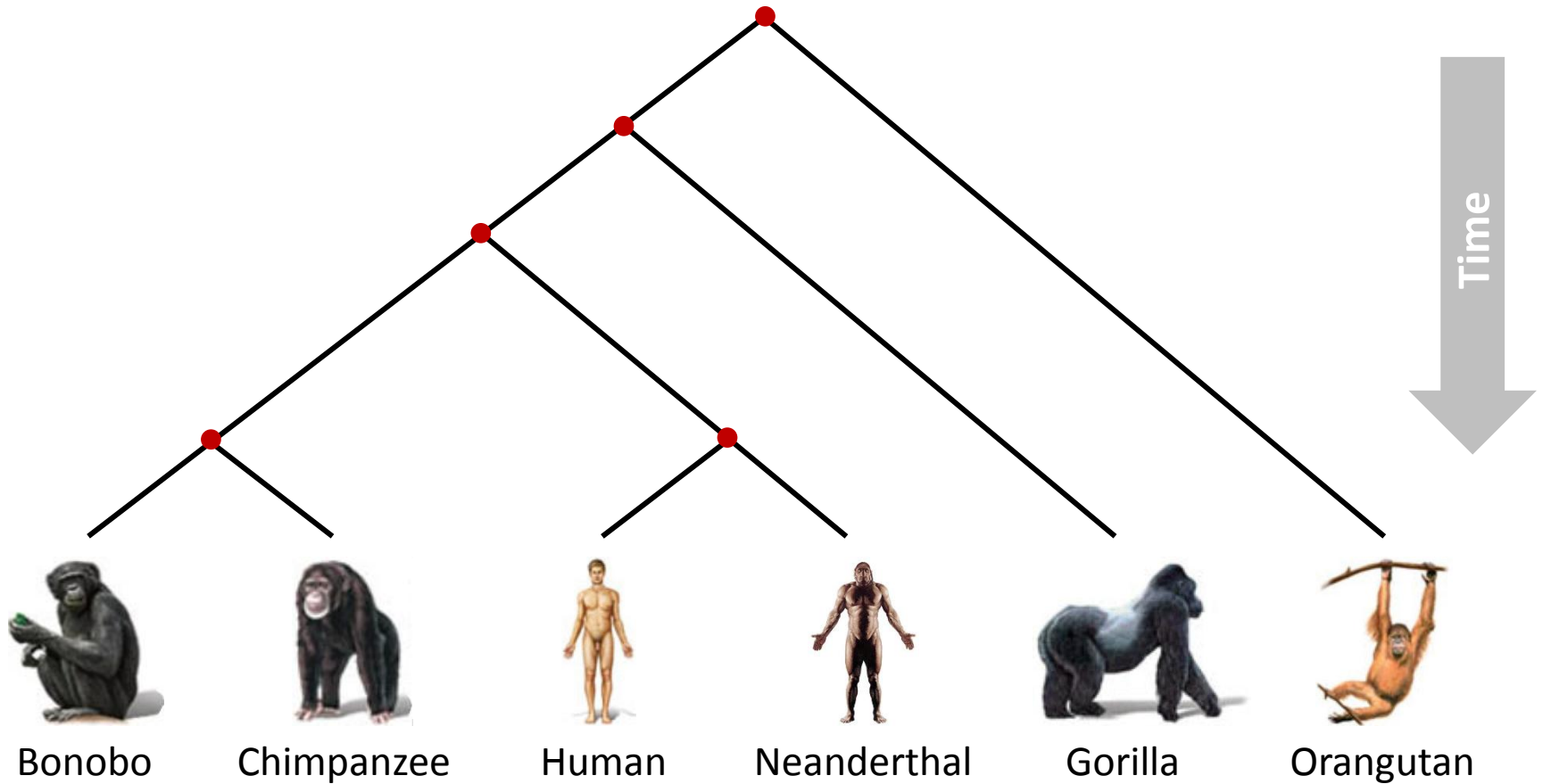
Rolf Fagerberg

University of Southern Denmark

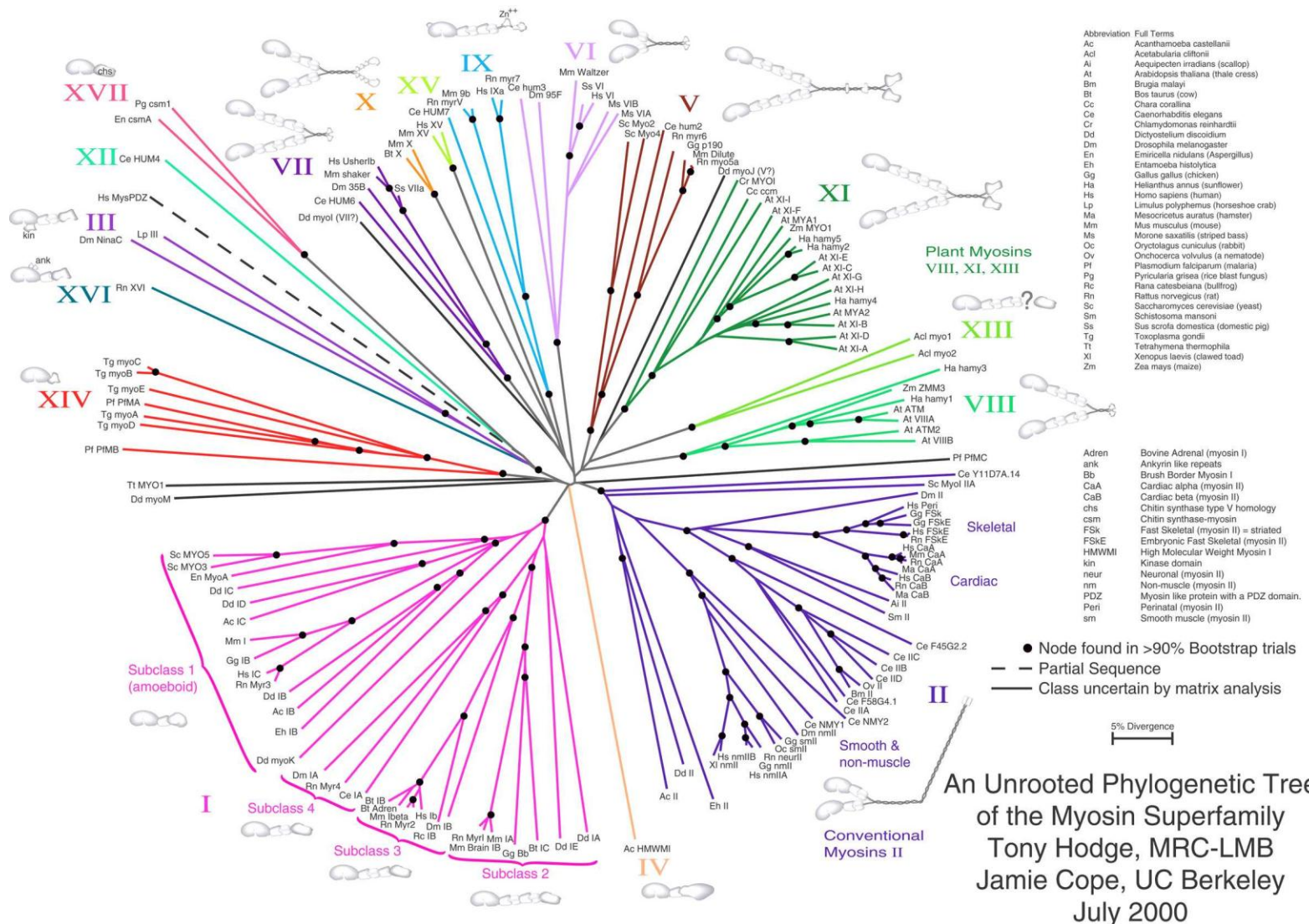
Thomas Mailund, Christian N. S. Pedersen, Andreas Sand

Aarhus University, Bioinformatics Research Center

Rooted Evolutionary Tree



Unrooted Evolutionary Tree



An Unrooted Phylogenetic Tree of the Myosin Superfamily
 Tony Hodge, MRC-LMB
 Jamie Cope, UC Berkeley
 July 2000

Dominant modern approach to study evolution is from DNA analysis

Constructing Evolutionary Trees – Binary or Arbitrary Degrees ?



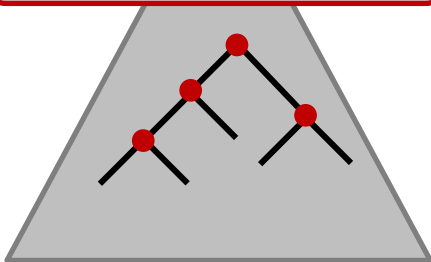
Distance matrix

	1	2	3	...	n
1					
2					
3					
⋮					
n					



Binary trees

(despite **no evidence** in distance data)



Neighbor Joining

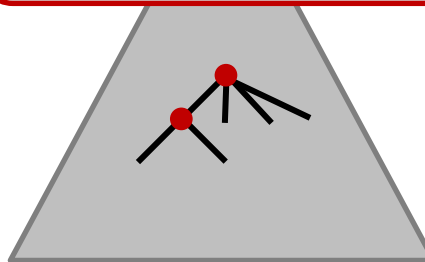
Saitou, Nei 1987

[$O(n^3)$ Saitou, Nei 1987]

....

Arbitrary degree

(**compromise** ; good support for all edges)



Refined Buneman Trees

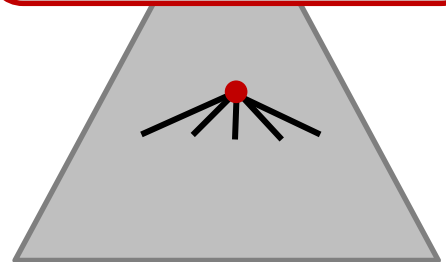
Moulton, Steel 1999

[$O(n^3)$ Brodal *et al.* 2003]

....

Arbitrary degrees

(strong support for all edges ; **few branches**)



Buneman Trees

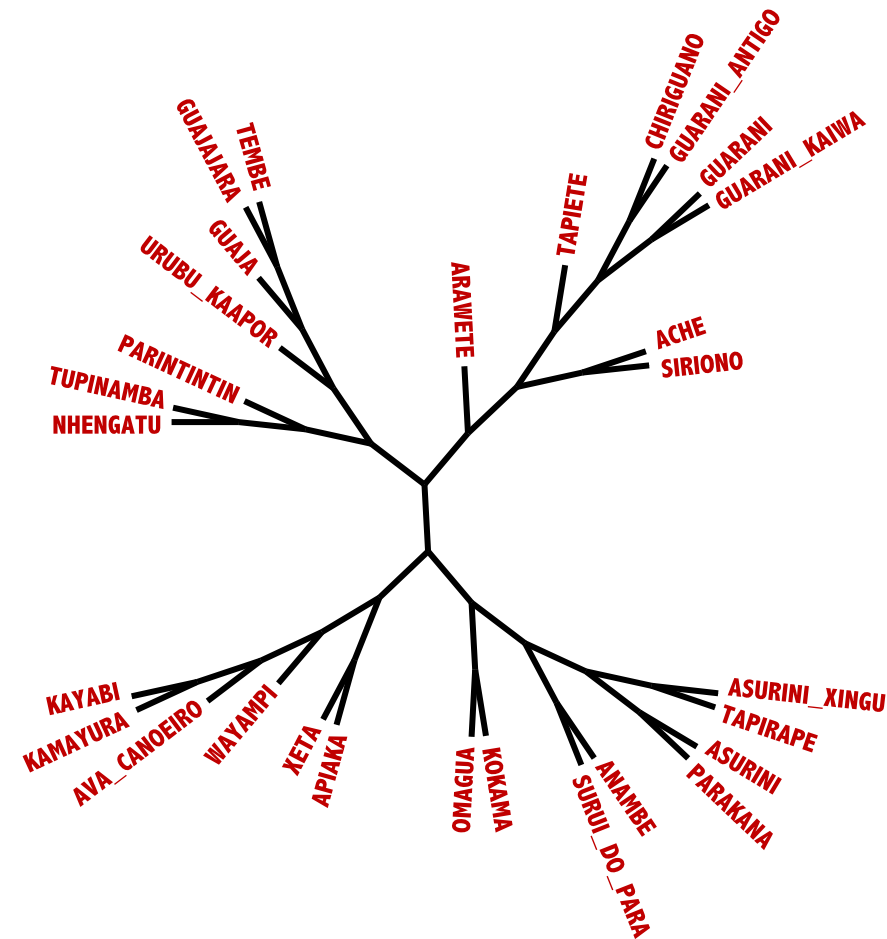
Buneman 1971

[$O(n^3)$ Berry, Bryan 1999]

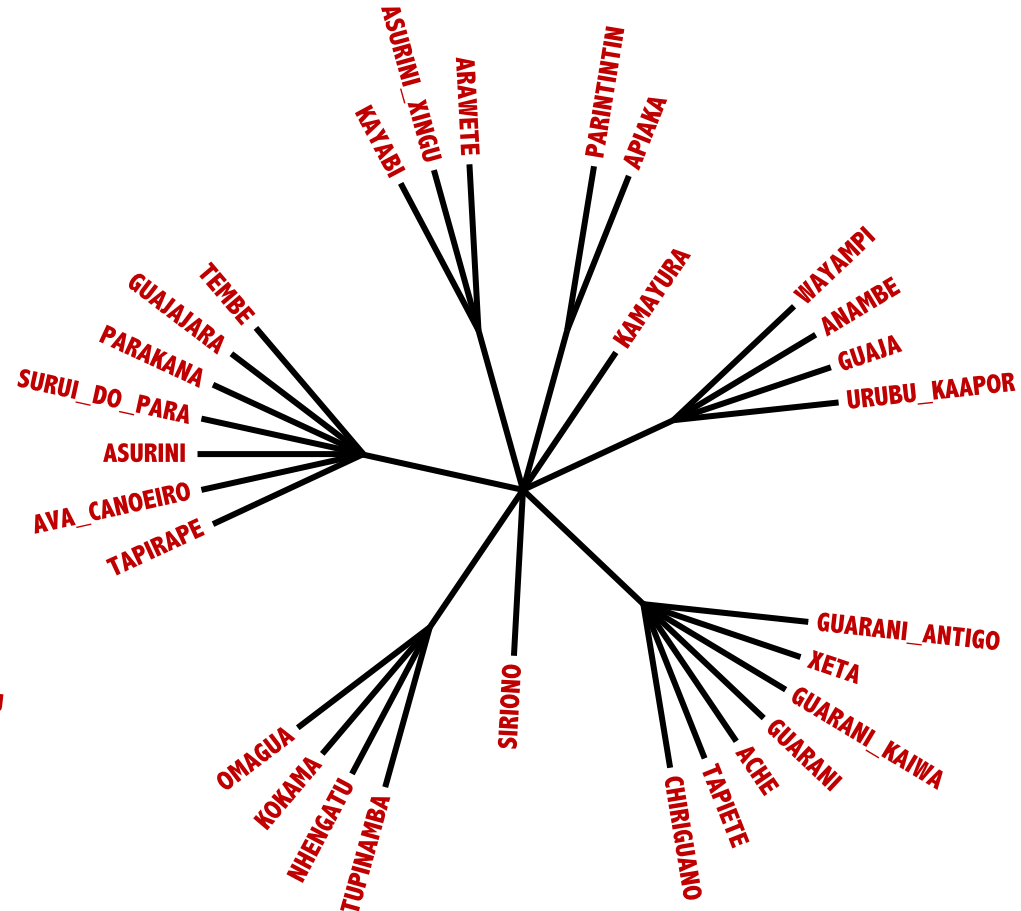
Data Analysis vs Expert Trees – Binary vs Arbitrary Degrees ?

Cultural Phylogenetics of the Tupi Language Family in Lowland South America.

R. S. Walker, S. Wichmann, T. Mailund, C. J. Atkisson. PLoS One. 7(4), 2012.

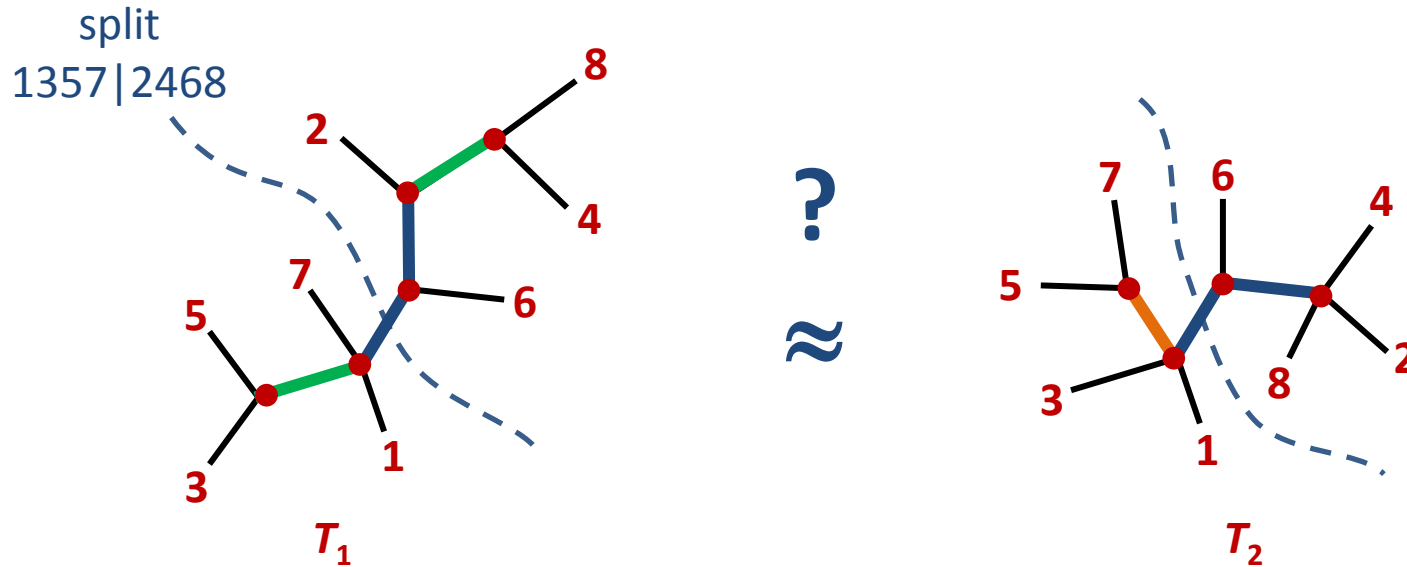


Neighbor Joining on linguistic data



Linguistic expert classification
(Aryon Rodrigues)

Evolutionary Tree Comparison



Common	Only T_1	Only T_2
1357 2468	35 124678	57 123468
13567 248	48 123567	

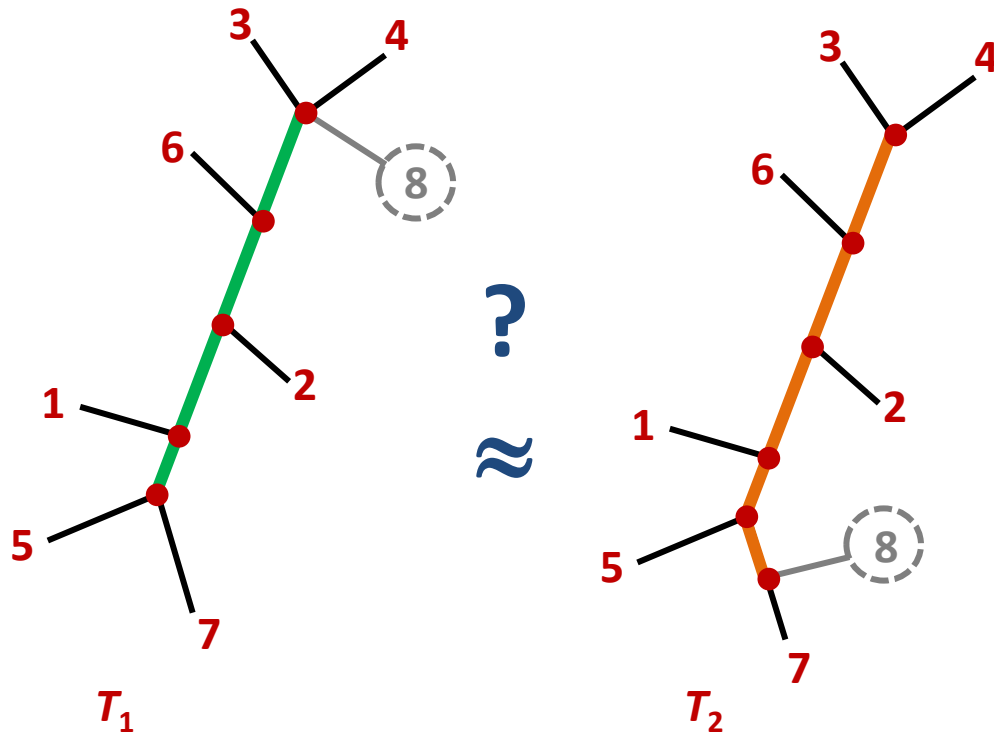
Robinson-Foulds distance = # non-common splits = **2** + **1** = **3**

D. F. Robinson and L. R. Foulds. Comparison of weighted labeled trees. In *Combinatorial mathematics, VI*, Lecture Notes in Mathematics, pages 119–126. Springer, 1979.

[Day 1985] $O(n)$ time algorithm using 2 x DFS + radix sort

Robinson-Foulds Distance (unrooted trees)

D. F. Robinson and L. R. Foulds. Comparison of weighted labeled trees. In *Combinatorial mathematics, VI*, Lecture Notes in Mathematics, pages 119–126. Springer, 1979.



Common	Only T_1	Only T_2
(none)	12567 348	125678 34
	1257 3468	12578 346
	157 23468	1578 2346
	57 123468	578 12346
		78 123456

$$\text{RF-dist}(T_1, T_2) = 4 + 5 = 9$$

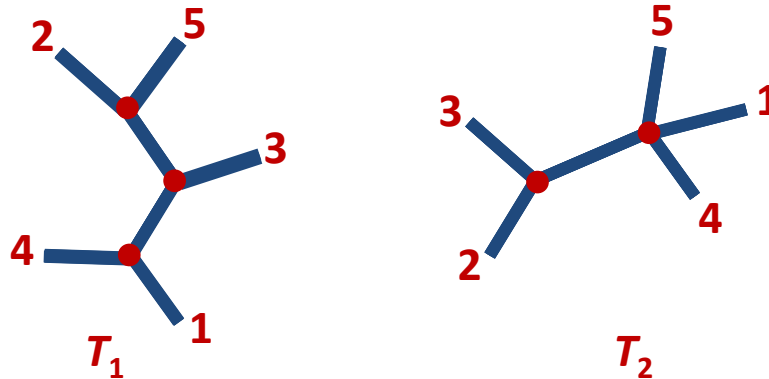
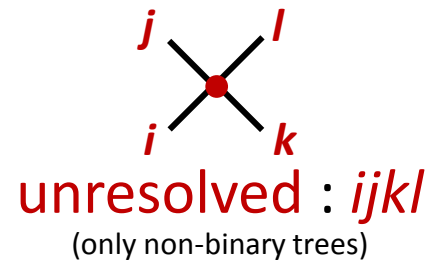
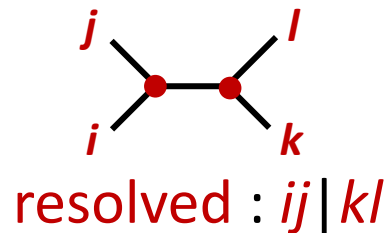
$$\text{RF-dist}(T_1 \setminus \{8\}, T_2 \setminus \{8\}) = 0$$

Robinson-Foulds very sensitive to outliers

Quartet Distance (unrooted trees)

G. Estabrook, F. McMorris, and C. Meacham. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, 34:193-200, 1985.

Consider all $\binom{n}{4}$ **quartets**, i.e. topologies of subsets of 4 leaves $\{i,j,k,l\}$



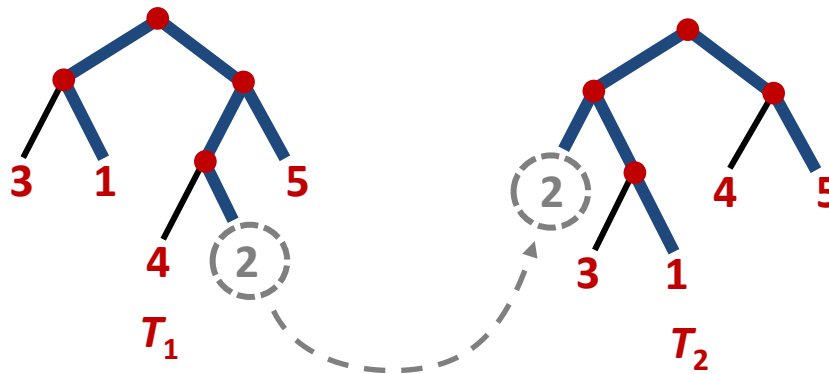
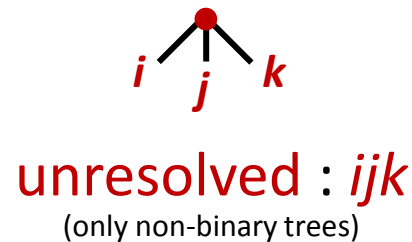
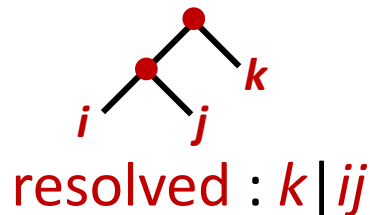
Quartet	T_1	T_2
$\{1,2,3,4\}$	14 23	14 23
$\{1,2,3,5\}$	13 25	15 23
$\{1,2,4,5\}$	14 25	1245
$\{1,3,4,5\}$	14 35	1345
$\{2,3,4,5\}$	25 34	23 45

$$\text{Quartet-dist}(T_1, T_2) = \binom{n}{4} - \text{\# common quartets} = 5 - 1 = 4$$

Triplet Distance (rooted trees)

D. E. Critchlow, D. K. Pearl, C. L. Qian: The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45(3):323-334, 1996.

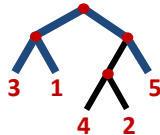
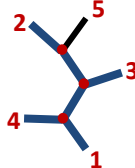
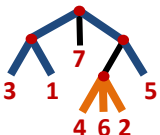
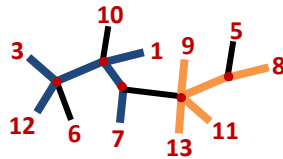
Consider all $\binom{n}{3}$ triplets, i.e. topologies of subsets of 3 leaves $\{i,j,k\}$



Triplet	T_1	T_2
$\{1,2,3\}$	2 13	2 13
$\{1,2,4\}$	1 24	4 12
$\{1,2,5\}$	1 25	5 12
$\{1,3,4\}$	4 13	4 13
$\{1,3,5\}$	5 13	5 13
$\{1,4,5\}$	1 45	1 45
$\{2,3,4\}$	3 24	4 23
$\{2,3,5\}$	3 25	5 23
$\{2,4,5\}$	5 24	2 45
$\{3,4,5\}$	3 45	3 45

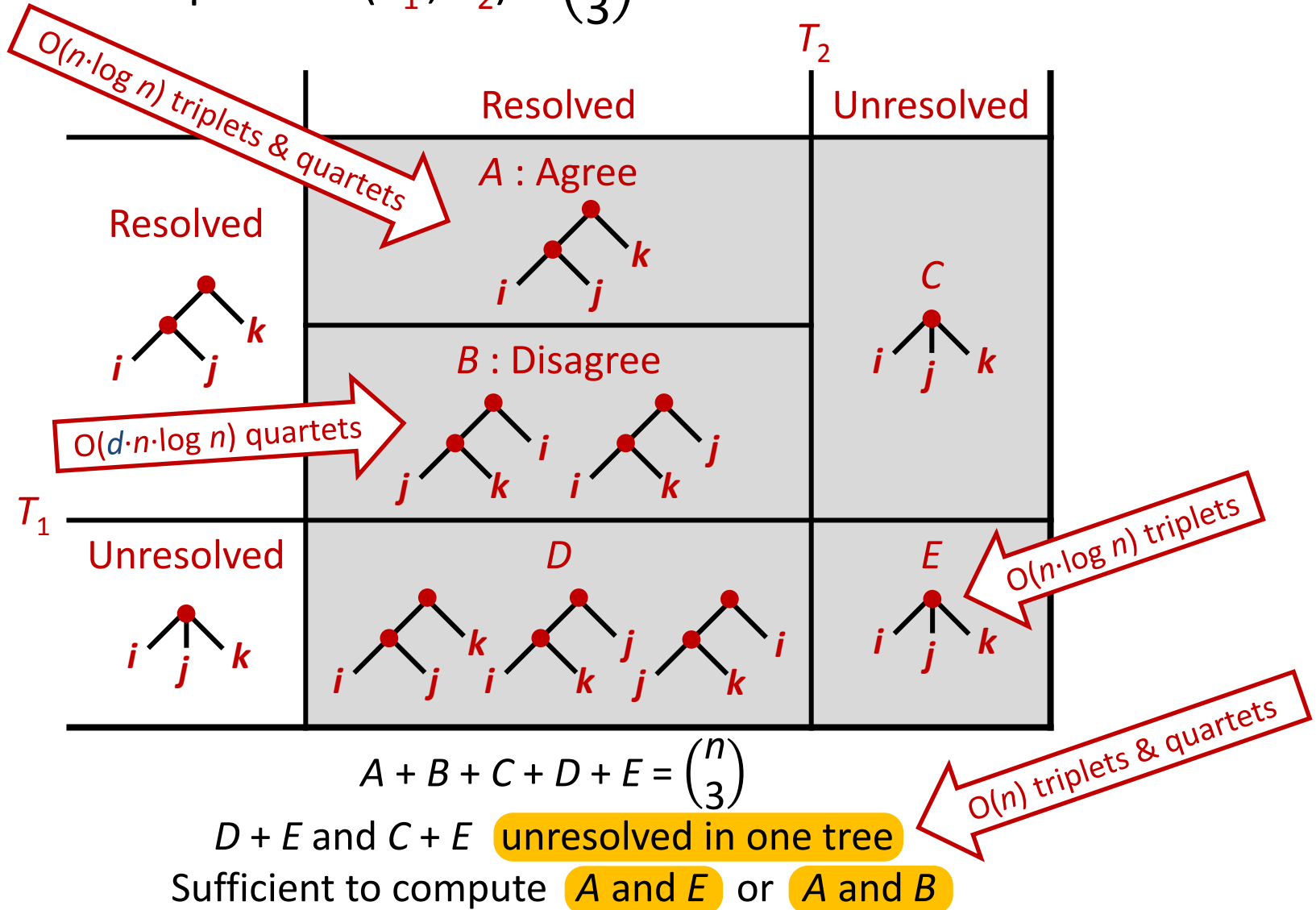
$$\text{Triplet-dist}(T_1, T_2) = \binom{n}{3} - \# \text{ common triplets} = 10 - 5 = 5$$

Computational Results

	Rooted Triplet distance	Unrooted Quartet distance
Binary	 $O(n^2)$ $O(n \cdot \log n)$ CPQ 1996 [SODA 2013]	 $O(n^3)$ $O(n^2)$ $O(n \cdot \log^2 n)$ $O(n \cdot \log n)$ D 1985 BTKL 2000 BFP 2001 BFP 2003
Degrees $\leq d$	 $O(n^2)$ $O(n \cdot \log n)$ BDF 2011 [SODA 2013]	 $O(d^9 \cdot n \cdot \log n)$ $O(n^{2.688})$ $O(d \cdot n \cdot \log n)$ SPMBF 2007 NKMP 2011 [SODA 2013]

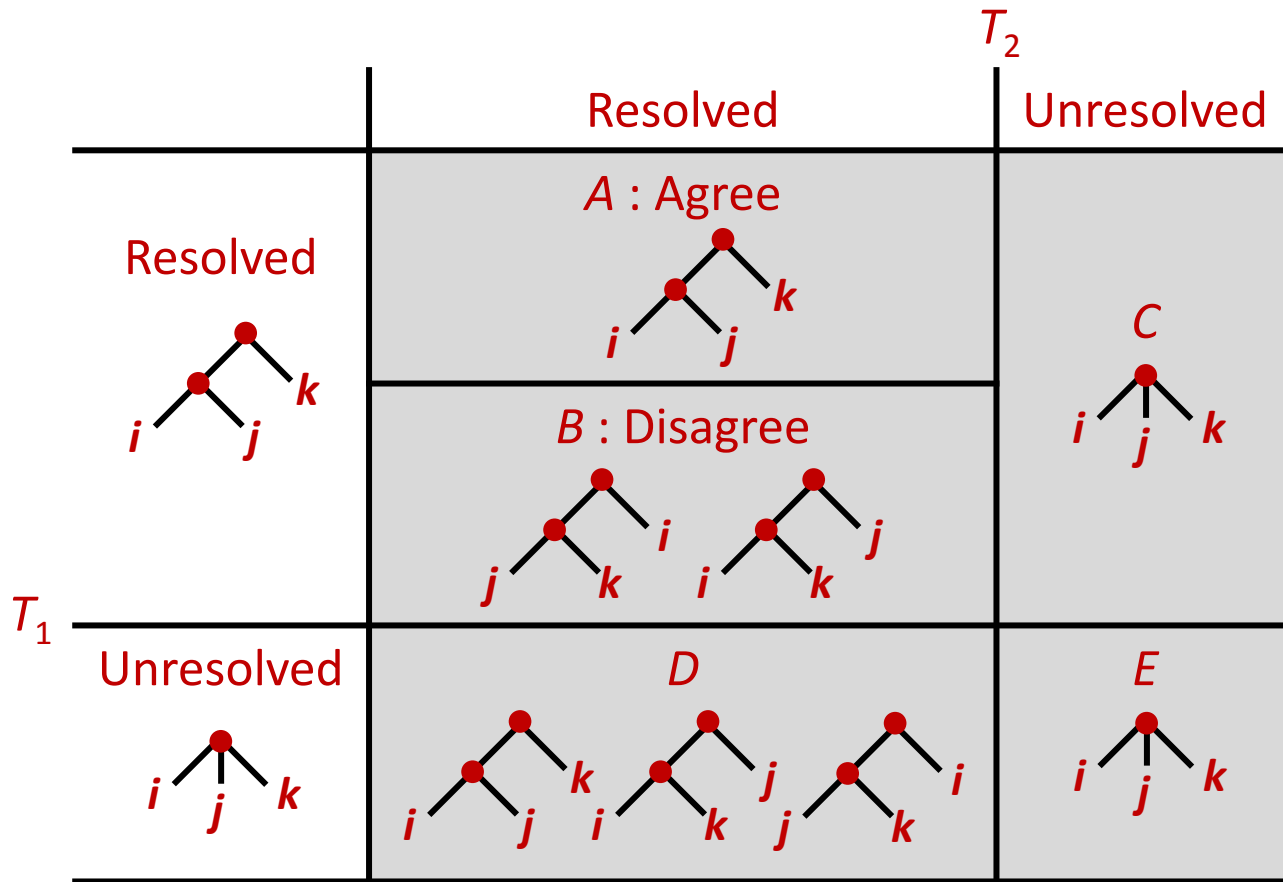
Distance Computation

$$\text{Triplet-dist}(T_1, T_2) = \binom{n}{3} - A - E = B + C + D$$



Parameterized Triplet & Quartet Distances

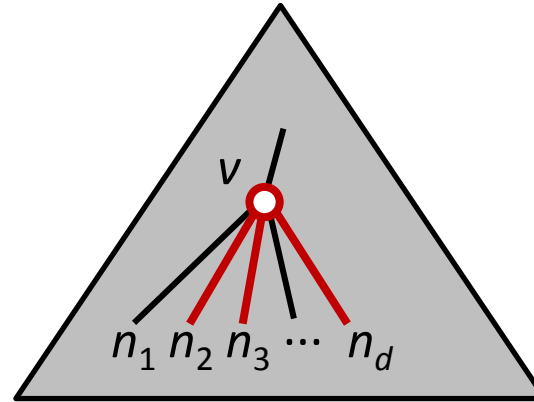
$$B + \alpha \cdot (C + D), \quad 0 \leq \alpha \leq 1$$



BDF 2011 $O(n^2)$ for triplet, NKMP 2011 $O(n^{2.688})$ for quartet
 [SODA 13] $O(n \cdot \log n)$ and $O(d \cdot n \cdot \log n)$, respectively

Counting Unresolved Triplets in One Tree

$$\sum_v \sum_{i < j < k} n_i \cdot n_j \cdot n_k$$

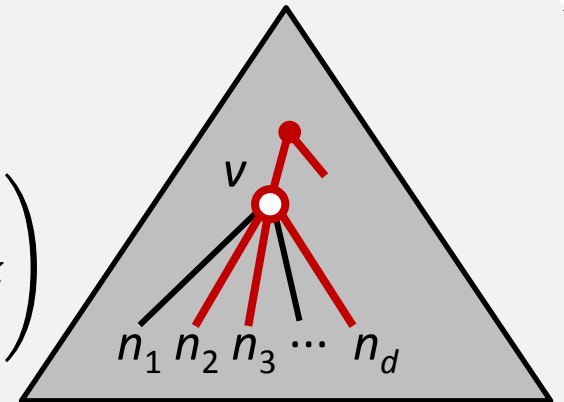


Triplet anchored at v

Computable in $O(n)$ time using DFS + dynamic programming

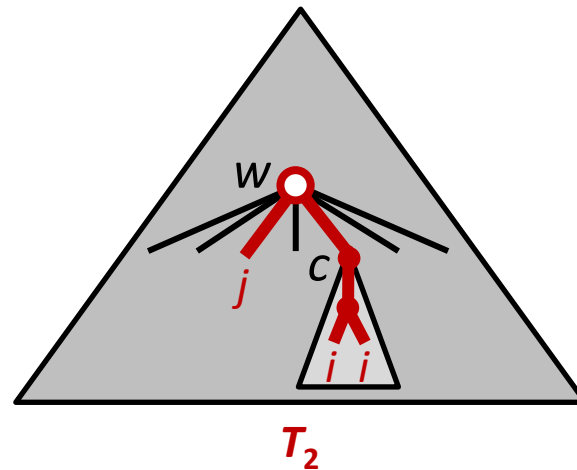
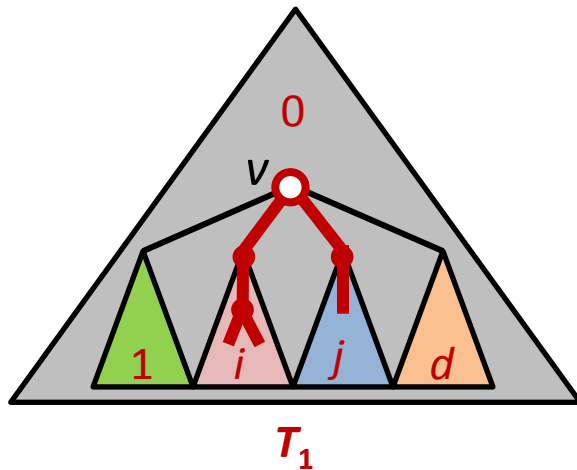
Quartets (root tree arbitrary)

$$\sum_v \left(\sum_{i < j < k < l} n_i \cdot n_j \cdot n_k \cdot n_l + \left(n - \sum_l n_l \right) \sum_{i < j < k} n_i \cdot n_j \cdot n_k \right)$$



Quartet anchored at v

Counting Agreeing Triplets (Basic Idea)

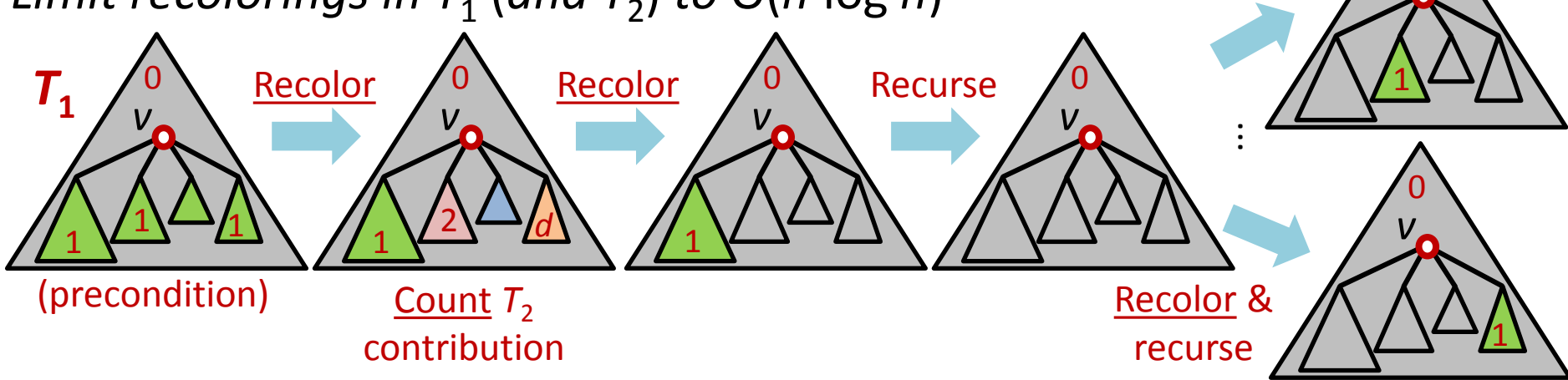


$$\sum_{v \in T_1} \sum_{w \in T_2} \sum_c \sum_{1 \leq i \leq d} \binom{n_i^c}{2} (n^w - n^c - n_i^w + n_i^c)$$

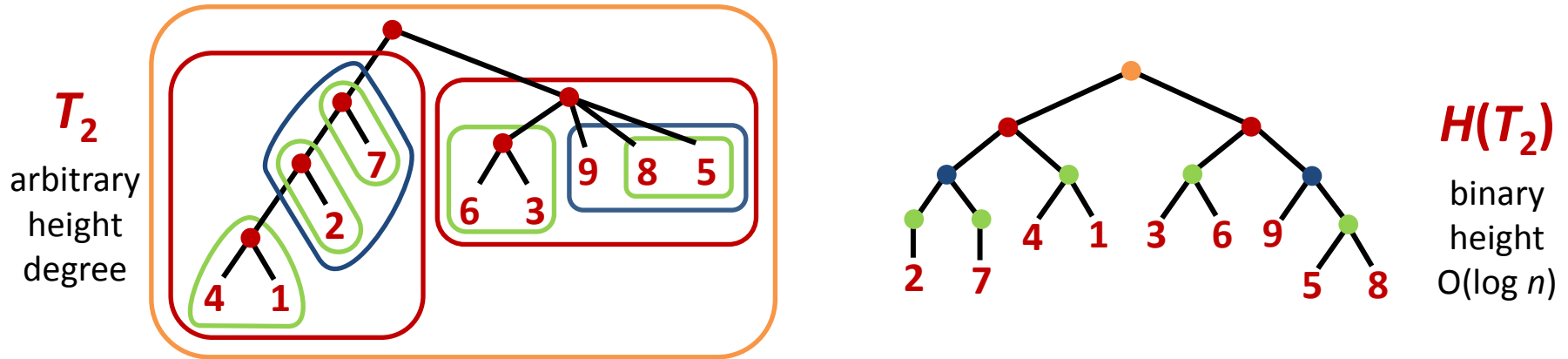
$\uparrow$
 $\sum_{1 \leq i \leq d} n_i^w$

Efficient Computation

Limit recolorings in T_1 (and T_2) to $O(n \cdot \log n)$



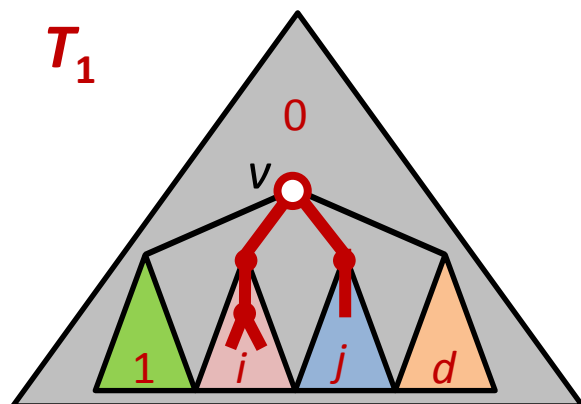
Reduce recoloring cost in T_2 to $O(n \cdot \log^2 n)$



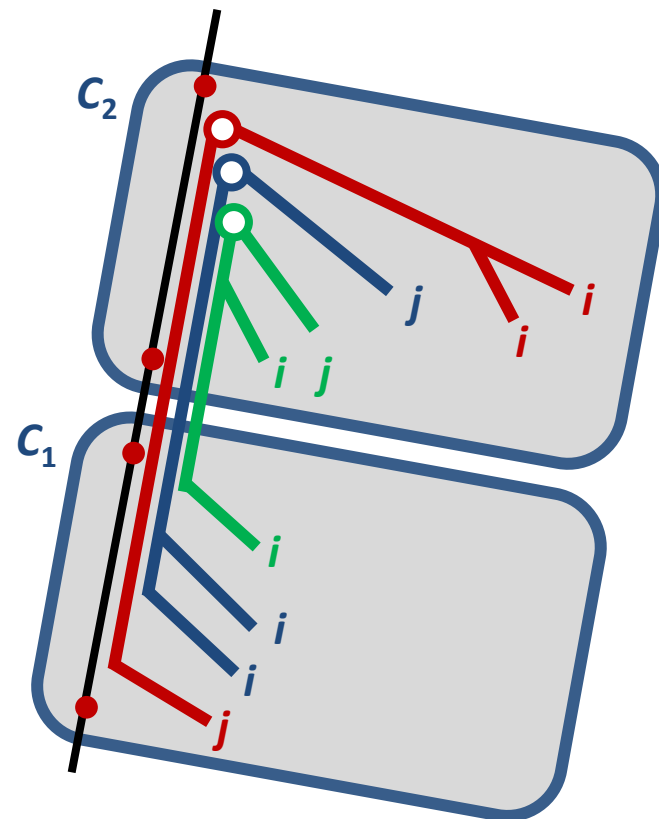
Reduce recoloring cost in T_2 from $O(n \cdot \log^2 n)$ to $O(n \cdot \log n)$

- Contract T_2 and reconstruct $H(T_2)$ during recursion

Counting Agreeing Triplets (II)



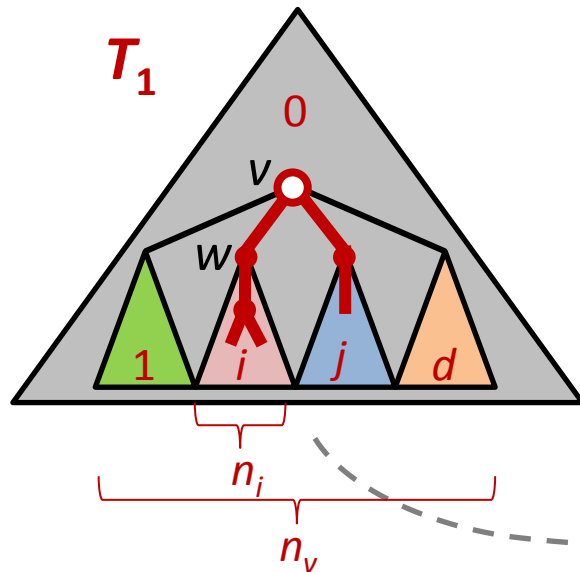
node in $H(T_2)$ =
component
composition in T_2



Contribution to agreeing triplets at node in $H(T_2)$

$$\sum_{1 \leq i \leq d} n_i^{C_1} \cdot n_{i \uparrow *}^{C_2} + \sum_{1 \leq i \leq d} \binom{n_i^{C_1}}{2} (n_*^{C_2} - n_i^{C_2}) + \sum_{1 \leq i \leq d} (n_*^{C_1} - n_i^{C_1}) n_{(ii)}^{C_2}$$

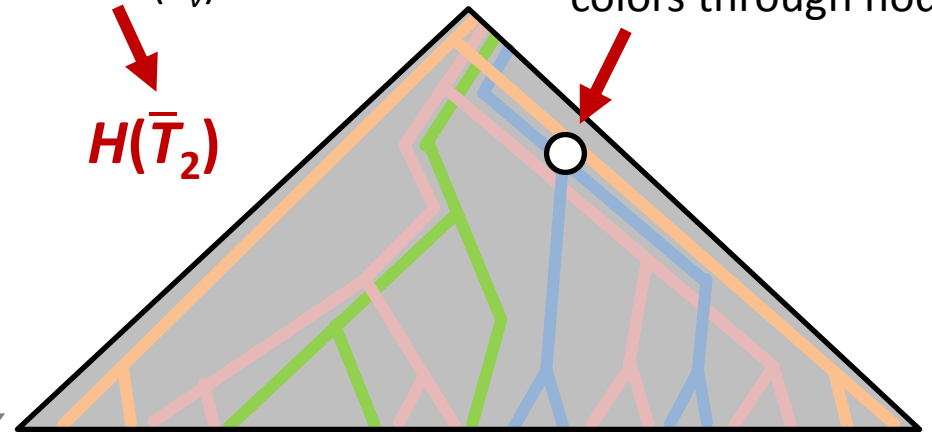
From $O(n \cdot \log^2 n)$ to $O(n \cdot \log n)$



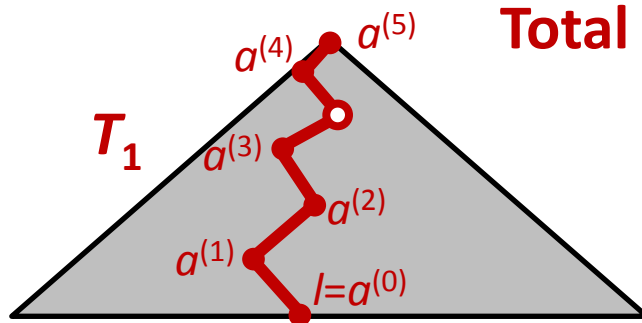
Compressed version
of T_2 of size $O(n_v)$

$H(\bar{T}_2)$

Update $O(1)$ counters for all
colors through node



Colored path lengths $\sum_{2 \leq i \leq d} \log \binom{|\bar{T}_2|}{n_i} = \sum_{2 \leq i \leq d} n_i \cdot \log \frac{n_v}{n_i}$

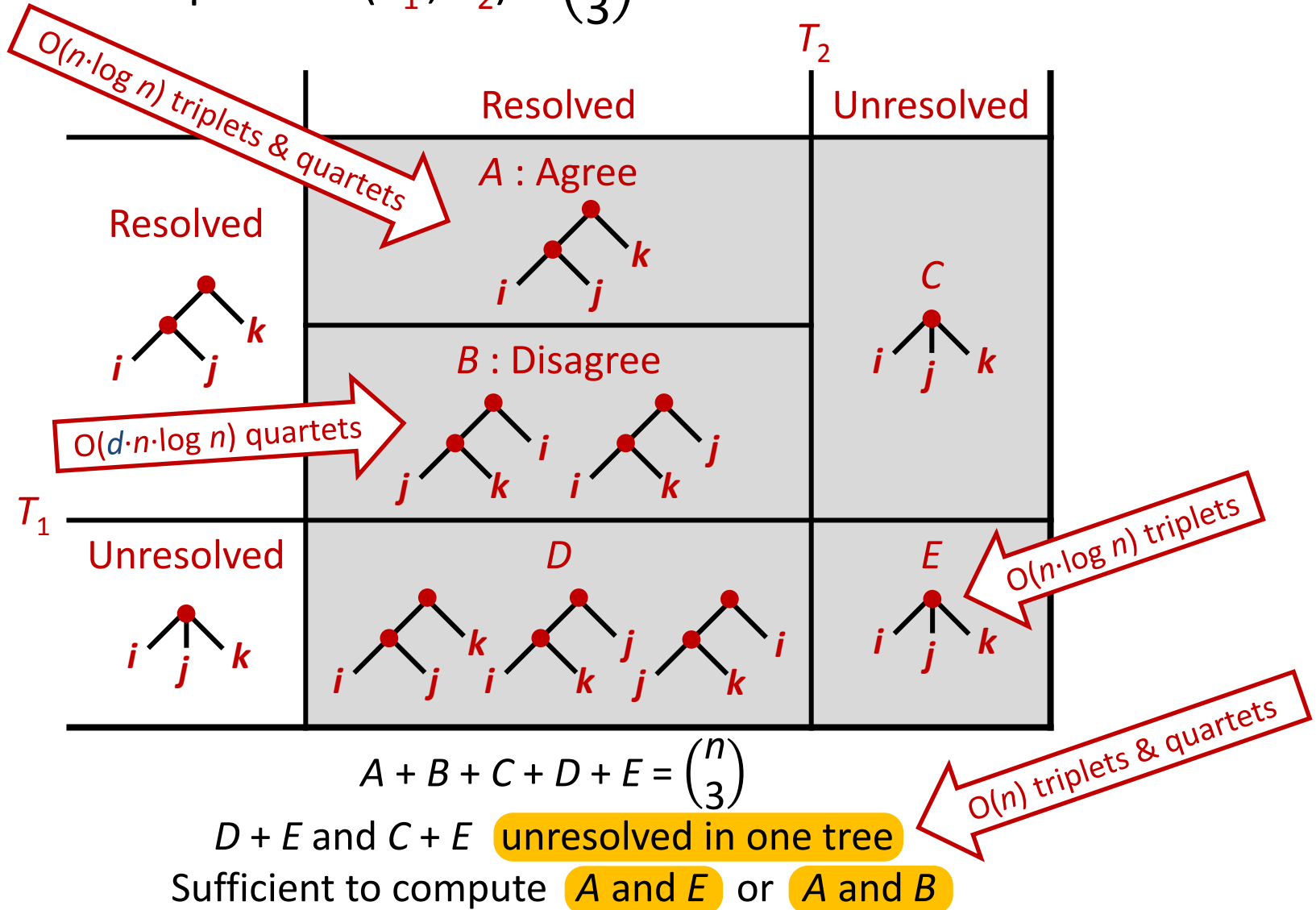


Total cost for updating counters

$$\sum_{\text{leaf } l \in T_1} \sum_{\substack{\text{ancestor } a^{(j)} \\ \text{not heavy child}}} \log \frac{n a^{(j+1)}}{n a^{(j)}} = n \cdot \log n$$

Distance Computation

$$\text{Triplet-dist}(T_1, T_2) = \binom{n}{3} - A - E = B + C + D$$



Summary

	Rooted Triplet distance	Unrooted Quartet distance
Binational 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 1 3 1 4 2 5 5 2 5 4 1 3 <		

d = maximal degree of any node in T_1 and T_2

$O(n \cdot \log n)$?