Triplet and Quartet Distances Between Trees of Arbitrary Degree

(paper to be presented at SODA'13)

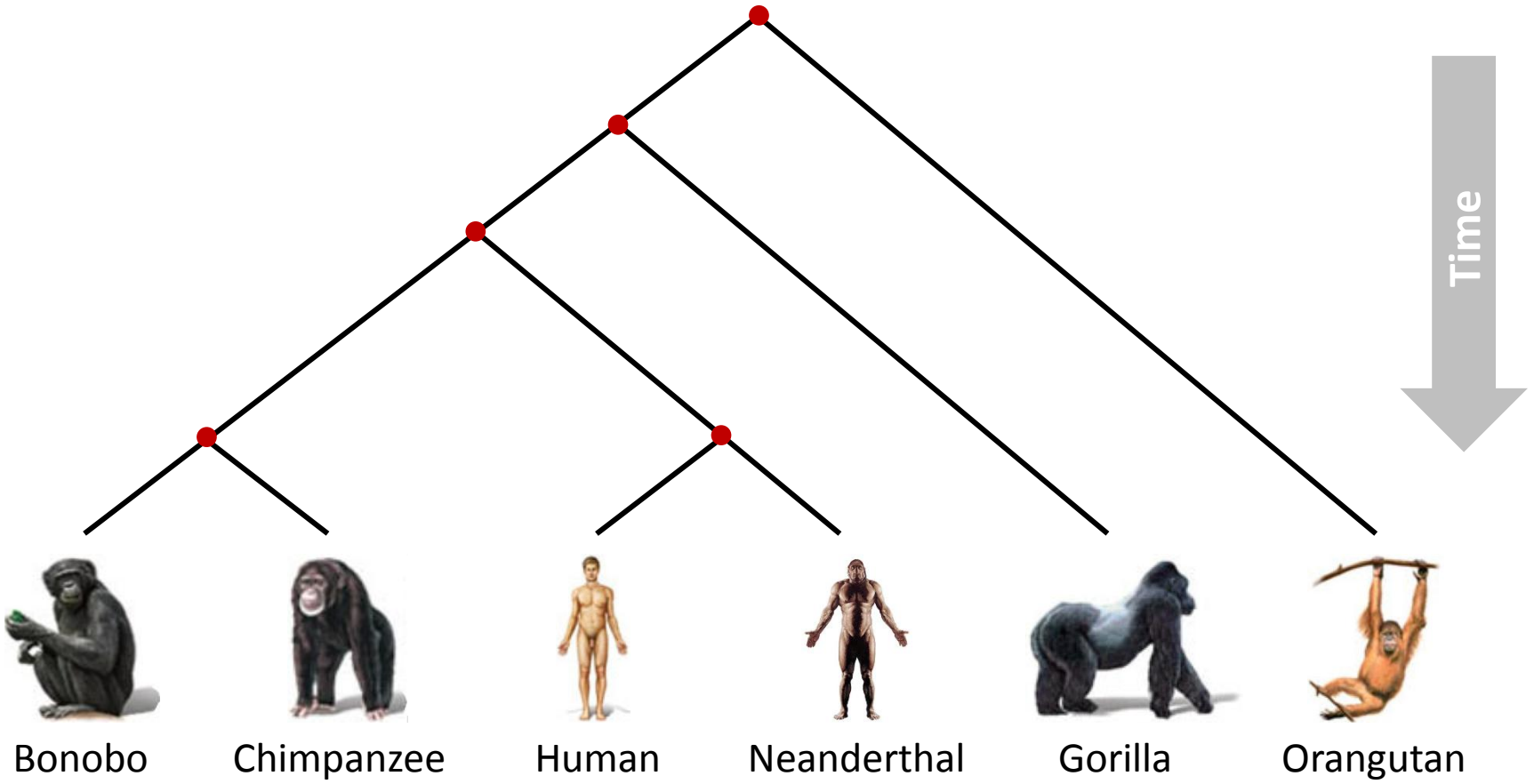
Gerth Stølting Brodal
Aarhus University

Rolf Fagerberg
University of Southern Denmark

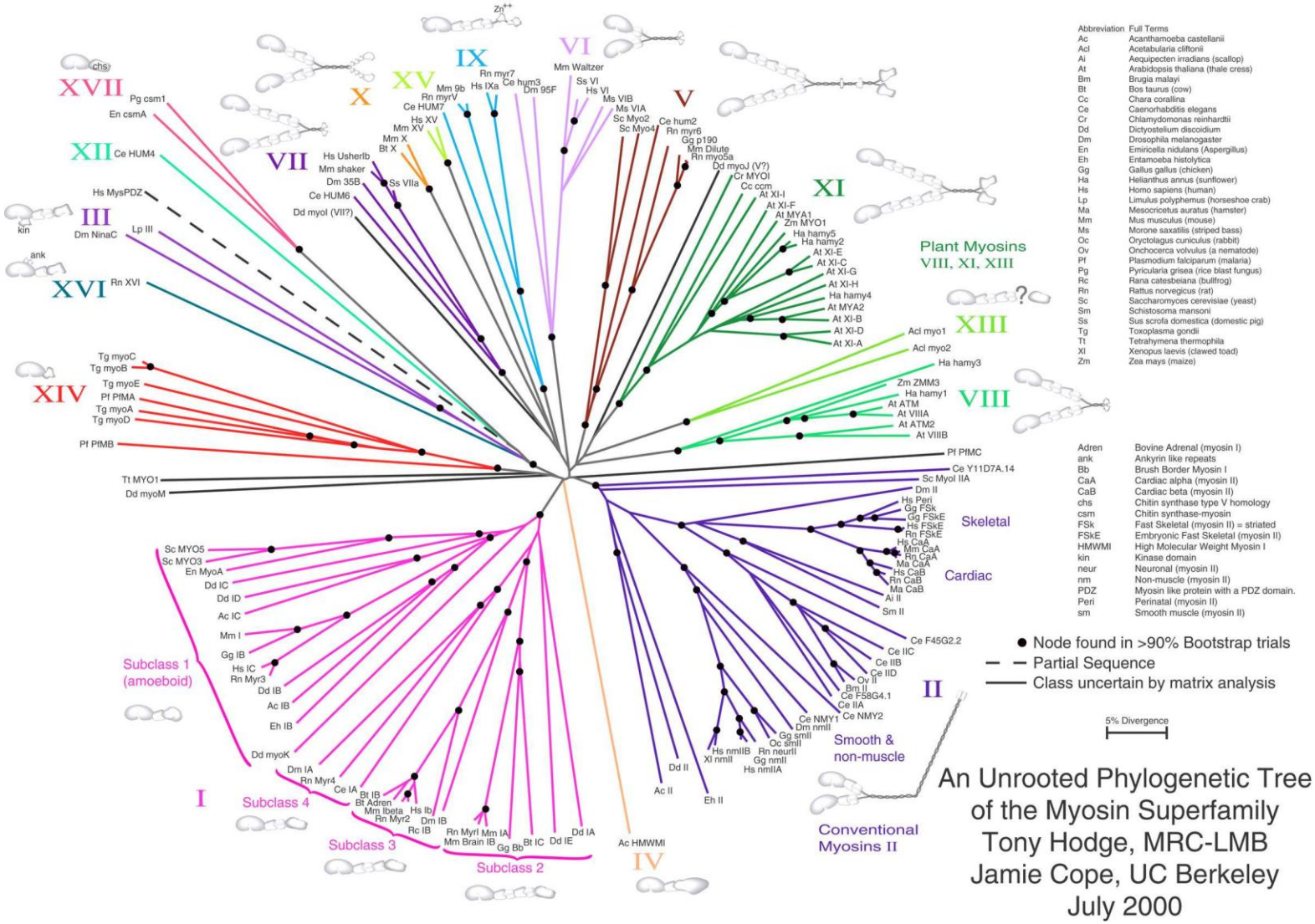
Thomas Mailund, Christian N. S. Pedersen, Andreas Sand
Aarhus University, Bioinformatics Research Center

ETH Zürich, Switzerland, 22 November 2012

Rooted Evolutionary Tree



Unrooted Evolutionary Tree

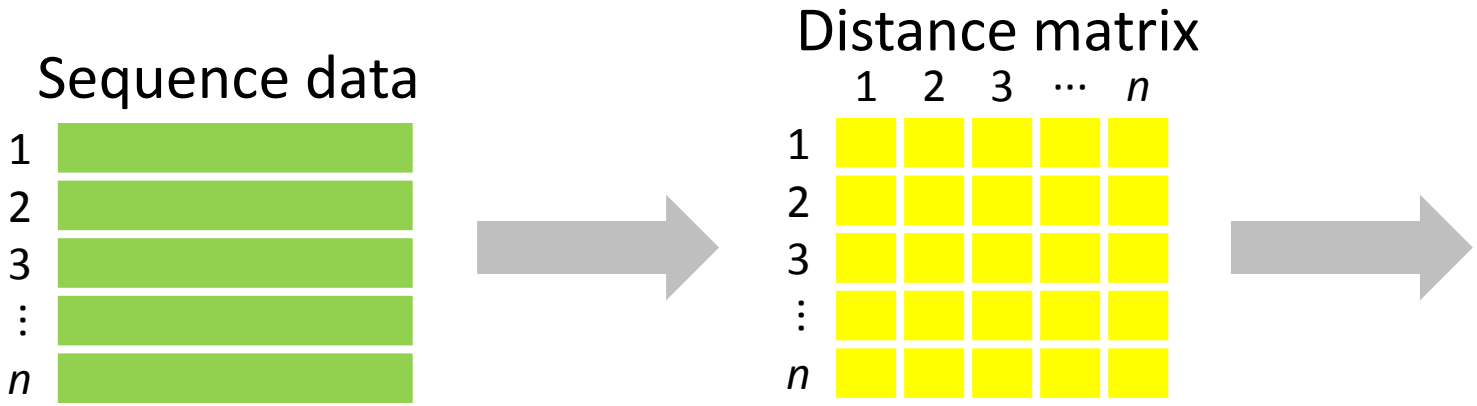


Abbreviation	Full Terms
Ac	Acanthamoeba castellanii
Acl	Acetabularia cliftonii
Ai	Aequipecten irradians (scallop)
At	Arabidopsis thaliana (thale cress)
Bm	Brugia malayi
Bt	Bos taurus (cow)
Cc	Chara corallina
Ce	Caenorhabditis elegans
Cr	Chlamydomonas reinhardtii
Dd	Dictyostelium discoidium
Dm	Drosophila melanogaster
En	Emiricella nidulans (Aspergillus)
Eh	Entamoeba histolytica
Gg	Gallus gallus (chicken)
Ha	Helianthus annus (sunflower)
Hs	Homo sapiens (human)
Lp	Limulus polyphemus (horseshoe crab)
Ma	Mesocricetus auratus (hamster)
Mm	Mus musculus (mouse)
Ms	Morone saxatilis (striped bass)
Oc	Oryctolagus cuniculus (rabbit)
Ov	Onchocerca volvulus (a nematode)
Pf	Plasmodium falciparum (malaria)
Pg	Pyricularia grisea (rice blast fungus)
Rc	Rana catesbeiana (bullfrog)
Rn	Rattus norvegicus (rat)
Sc	Saccharomyces cerevisiae (yeast)
Sm	Schistosoma mansoni
Ss	Sus scrofa domestica (domestic pig)
Tg	Toxoplasma gondii
Tt	Tetrahymena thermophila
Xl	Xenopus laevis (clawed toad)
Zm	Zea mays (maize)

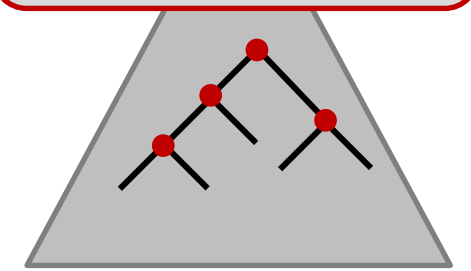
Adren	Bovine Adrenal (myosin I)
ank	Ankyrin like repeats
Bb	Brush Border Myosin I
CaA	Cardiac alpha (myosin II)
CaB	Cardiac beta (myosin II)
chs	Chitin synthase type V homology
csm	Chitin synthase-myosin
FSk	Fast Skeletal (myosin II) = striated
FSkE	Embryonic Fast Skeletal (myosin II)
HMWMI	High Molecular Weight Myosin I
kin	Kinase domain
neur	Neuronal (myosin II)
nm	Non-muscle (myosin II)
PDZ	Myosin like protein with a PDZ domain.
Peri	Perinatal (myosin II)
sm	Smooth muscle (myosin II)

Dominant modern approach to study evolution is from DNA analysis

Constructing Evolutionary Trees – Binary or Arbitrary Degrees ?



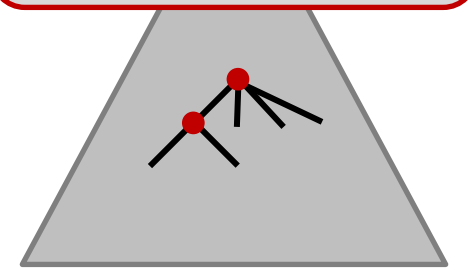
Binary trees
(despite no evidence in distance data)



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

....

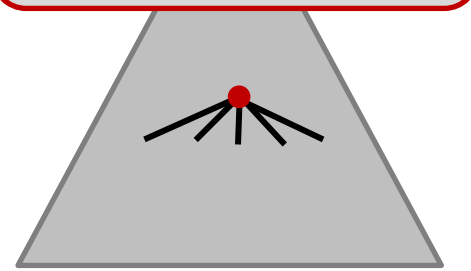
Arbitrary degree
(compromise ; good support for all edges)



Refined Buneman Trees
Moulton, Steel 1999
[O(n^3) Brodal *et al*. 2003]

....

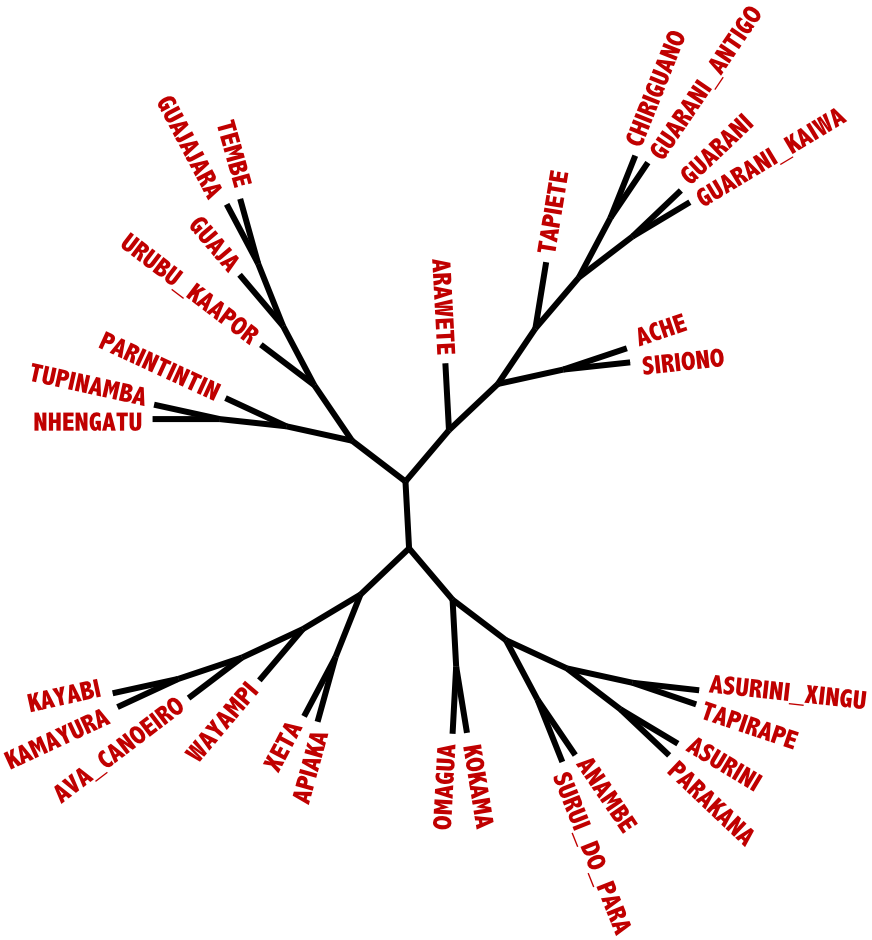
Arbitrary degrees
(strong support for all edges ; few branches)



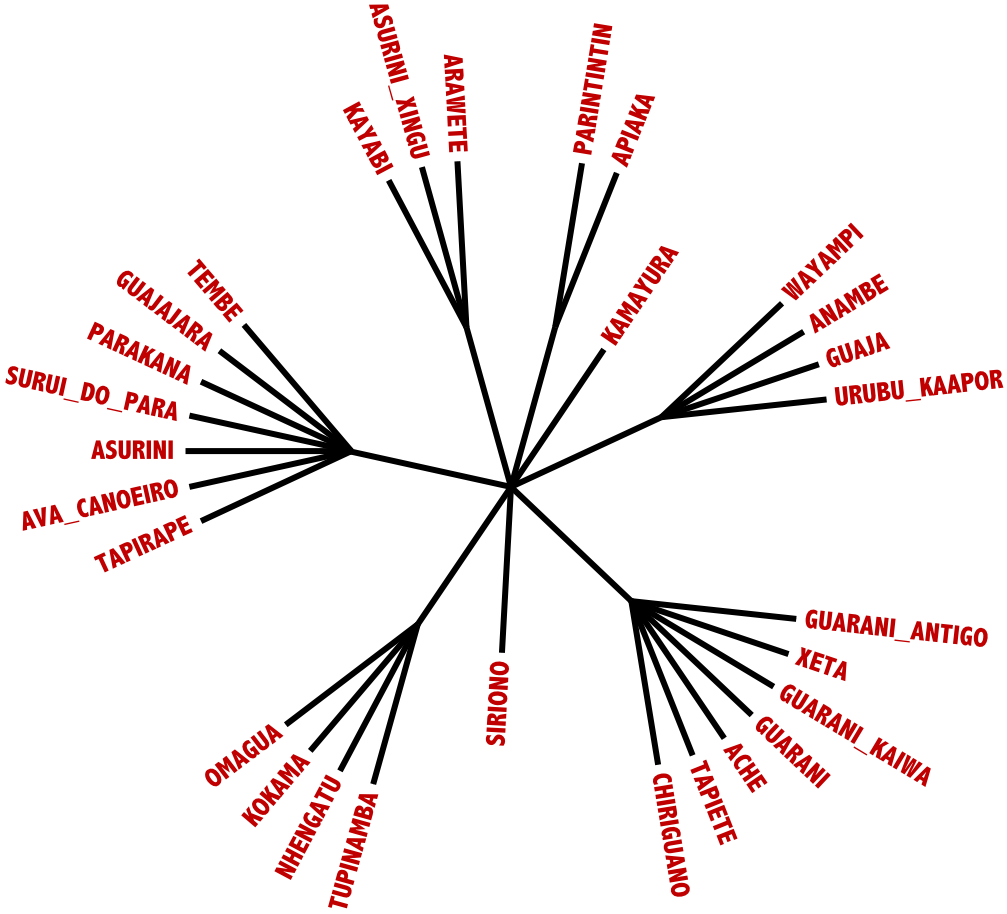
Buneman Trees
Buneman 1971
[O(n^3) Berry, Bryan 1999]

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.

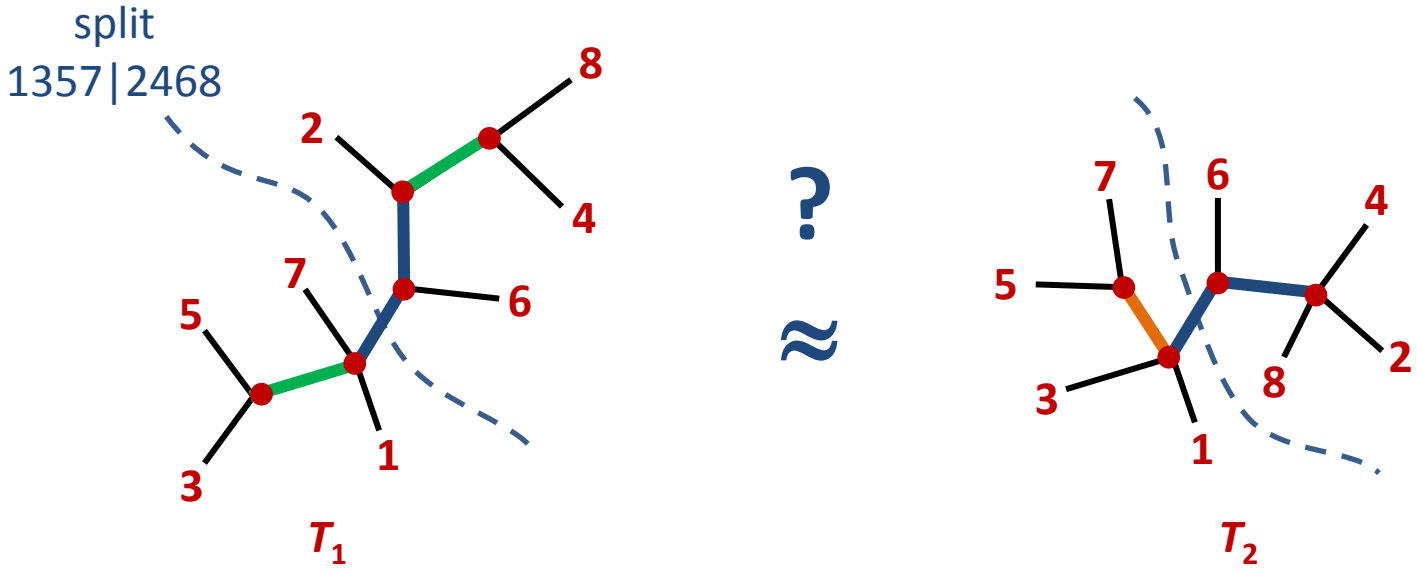


Neighbor Joining on linguistic data



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	

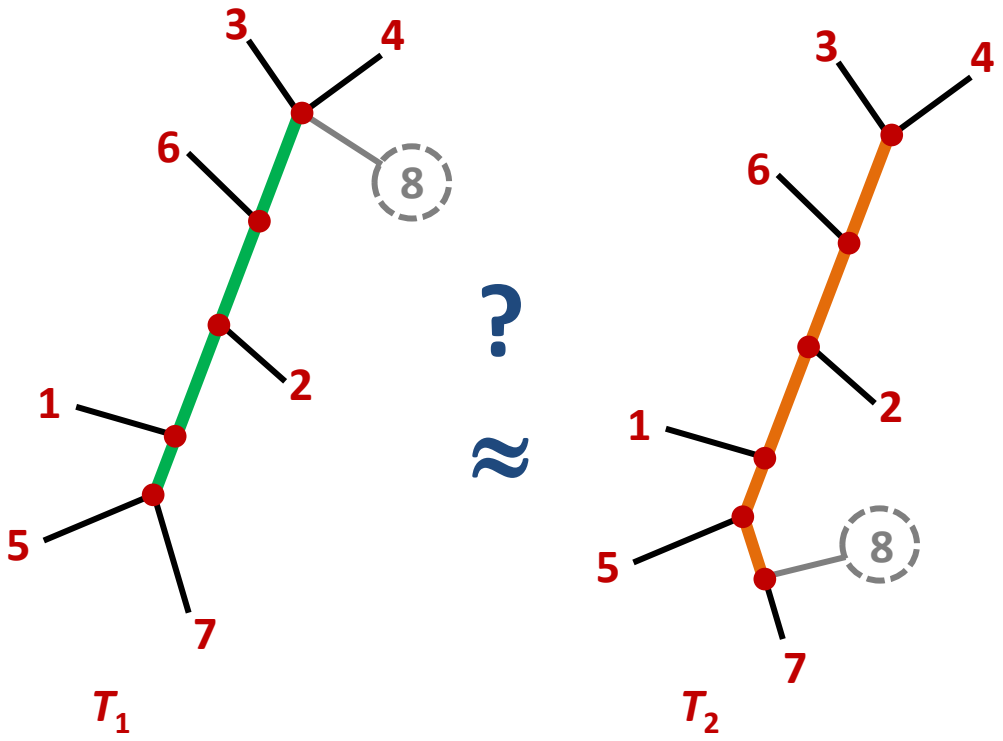
Robinson-Foulds distance = # non-common splits = **2** + **1** = **3**

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

RF-dist(T_1 , T_2) = **4** + **5** = **9**

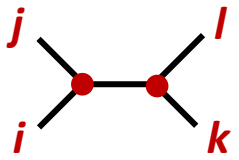
RF-dist(T_1\{8} , T_2\{8}) = **0**

Robinson-Foulds very sensitive to outliers

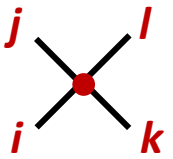
Quartet Distance (unrooted trees)

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 $\{i,j,k,l\}$

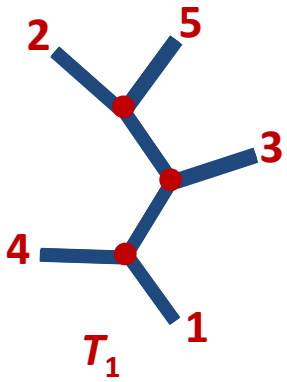


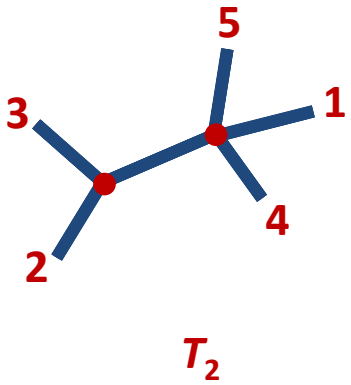
resolved : $ij|kl$



unresolved : $ijkl$

(only non-binary trees)





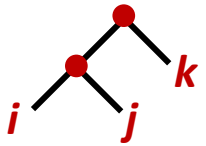
Quartet	T_1	T_2
{1,2,3,4}	14|23	14|23
{1,2,3,5}	13|25	15|23
{1,2,4,5}	14|25	1245
{1,3,4,5}	14|35	1345
{2,3,4,5}	25|34	23|45

Quartet-dist(T_1 , T_2) = $\binom{n}{4}$ - # 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.

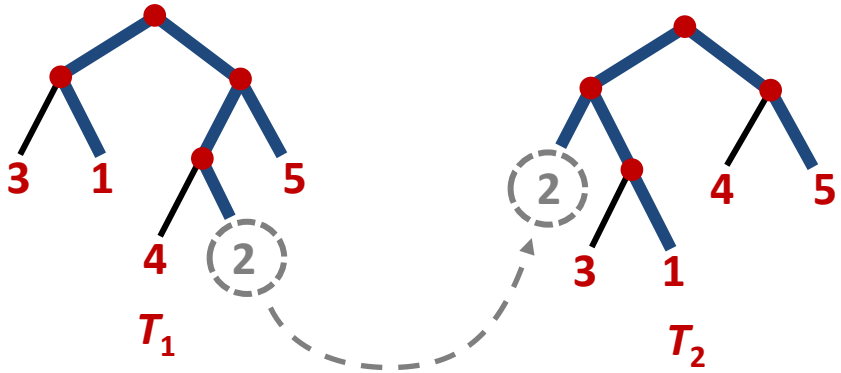
Consider all $\binom{n}{3}$ triplets, i.e. topologies of subsets of 3 leaves $\{i,j,k\}$



resolved : $k|ij$

unresolved : ijk

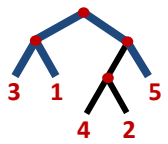
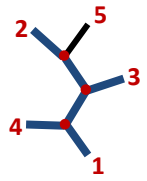
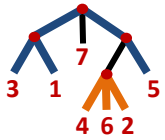
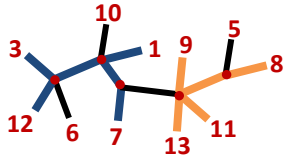
(only non-binary trees)



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

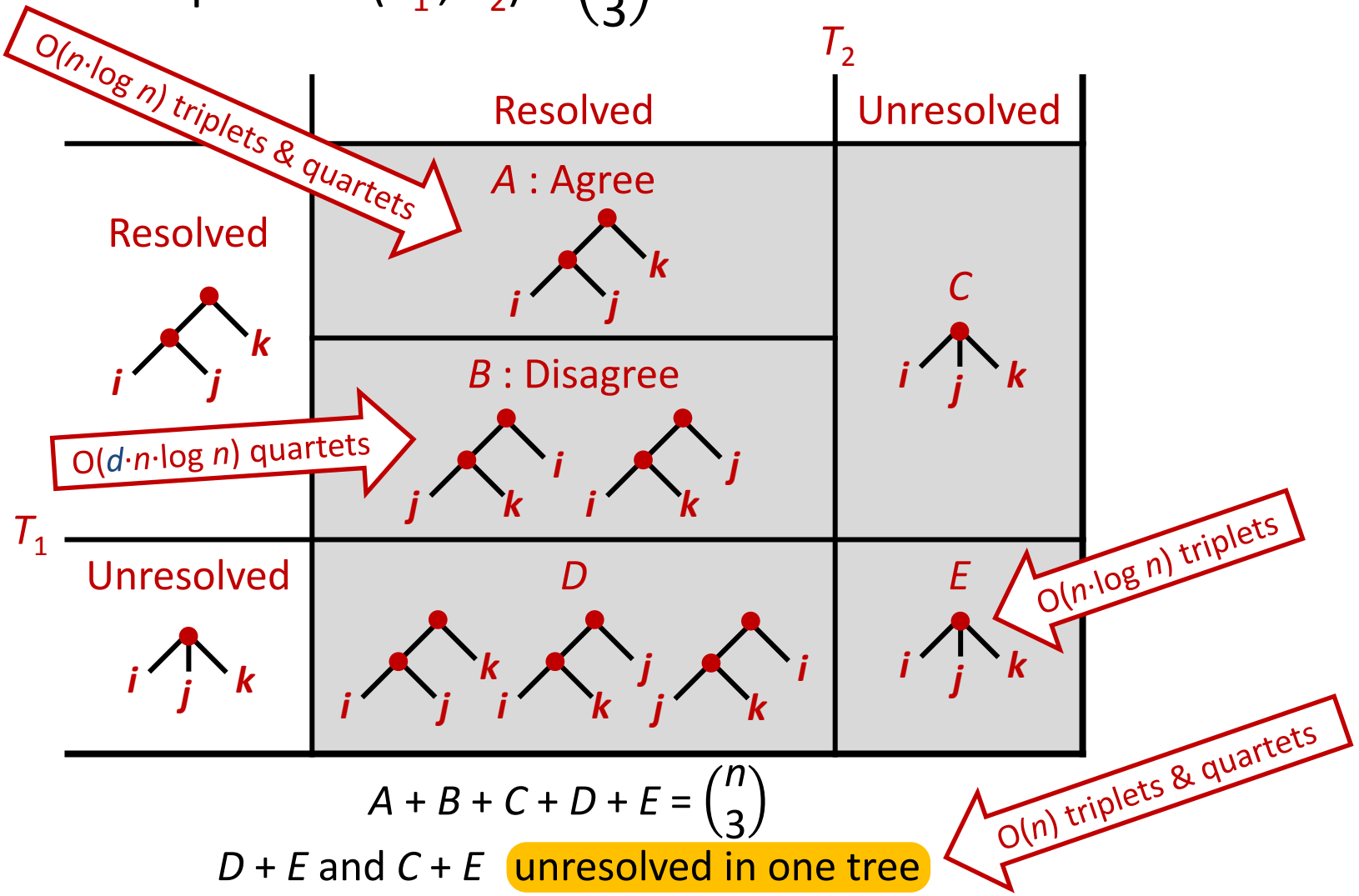
Triplet-dist(T_1 , T_2) = $\binom{n}{3}$ - # common triplets = **10** - **5** = **5**

Computational Results

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

Distance Computation

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



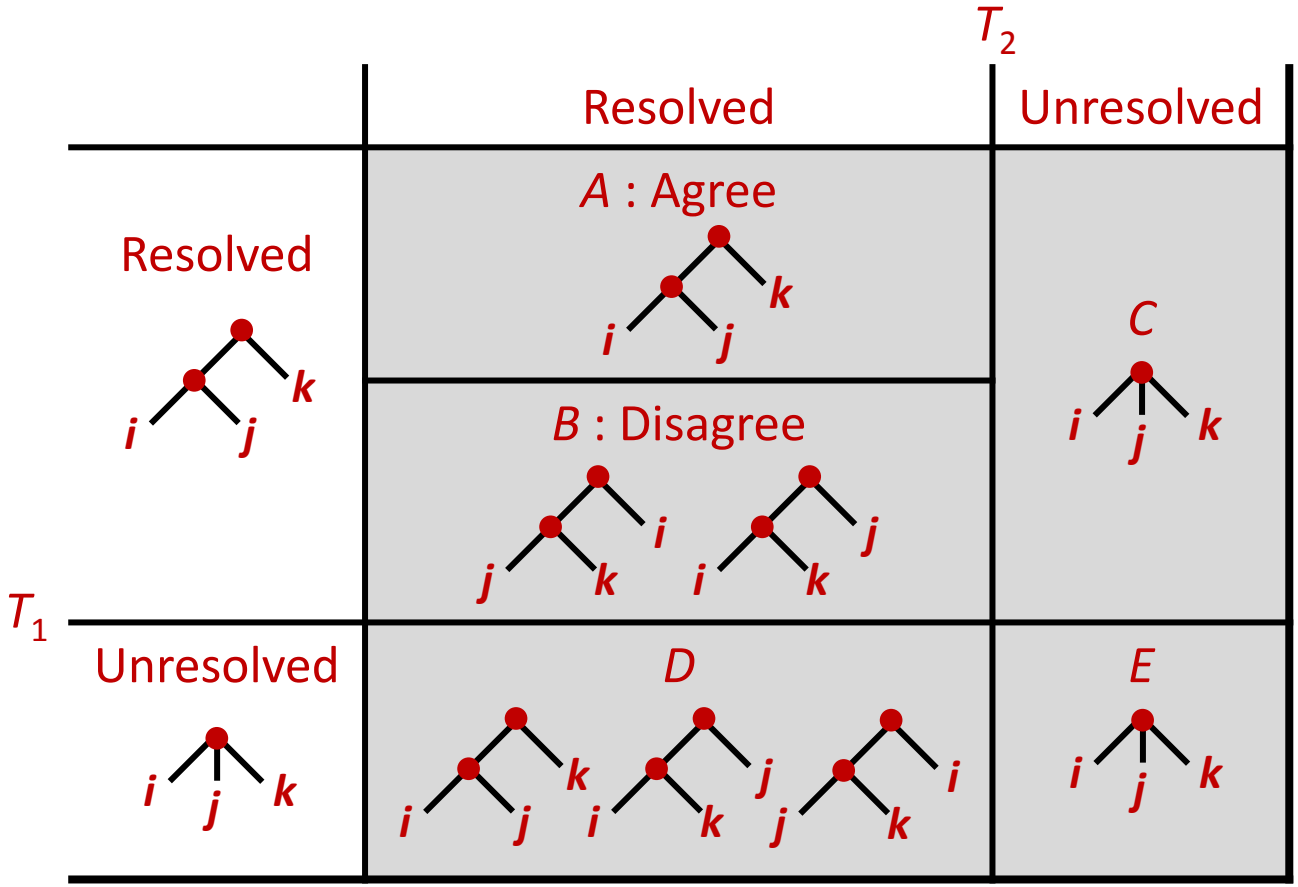
$A + B + C + D + E = \binom{n}{3}$

$D + E$ and $C + E$ unresolved in one tree

Sufficient to compute A and E or A and B

Parameterized Triplet & Quartet Distances

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

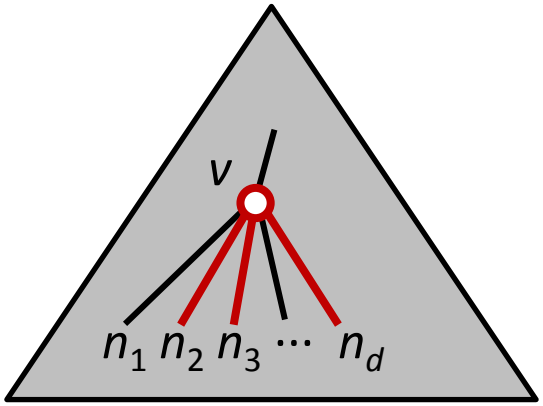


BDF 2011 $O(n^2)$ for triplet, NKMP 2011 $O(n^{2.688})$ for quartet

[SODA 13] $O(n \cdot \log n)$ and $O(d \cdot n \cdot \log n)$, respectively

Counting Unresolved Triplets in One Tree

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

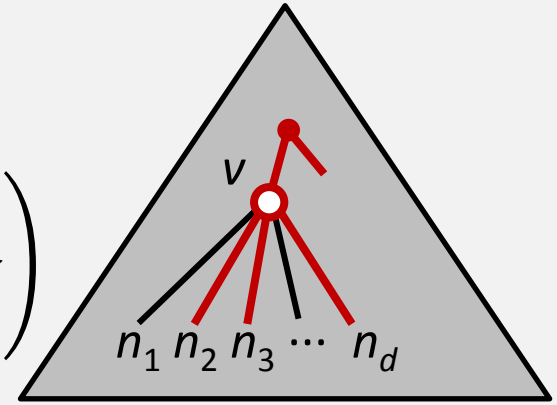


Triplet anchored at *v*

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

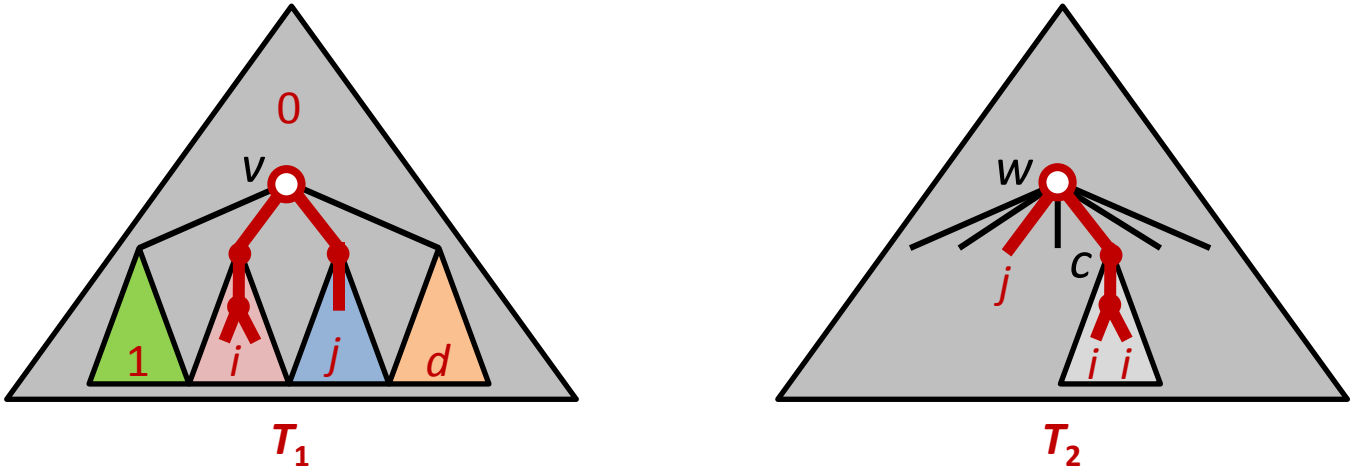
Quartets (root tree arbitrary)

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



Quartet anchored at *v*

Counting Agreeing Triplets (Basic Idea)

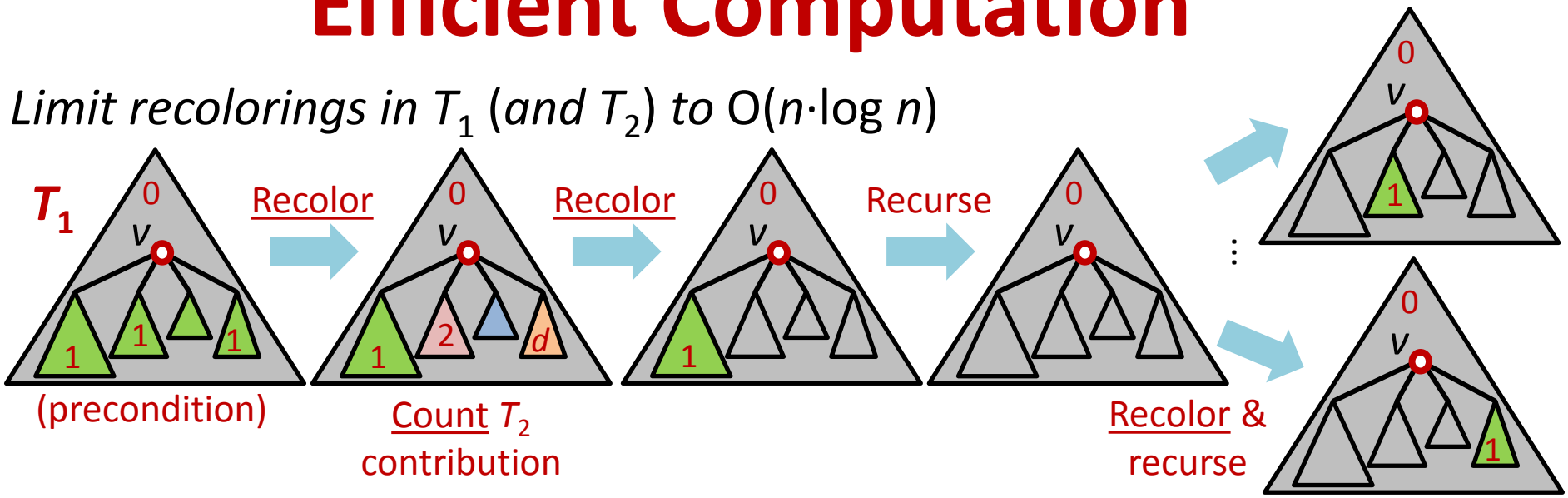


$$\sum_{v \in T_1} \sum_{w \in T_2} \sum_{c} \sum_{1 \le i \le d} \binom{n_i{}^c}{2} \left(n^w - n^c - n_i{}^w + n_i{}^c\right)$$

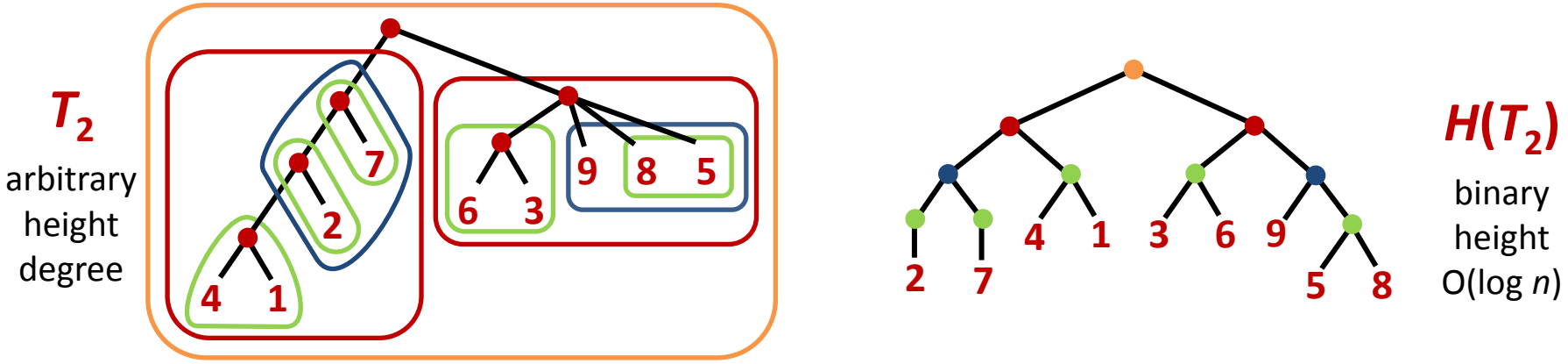
$$\sum_{1 \le i \le d} n_i{}^w$$

Efficient Computation

Limit recolorings in T_1 (and T_2) to O(n·log n)



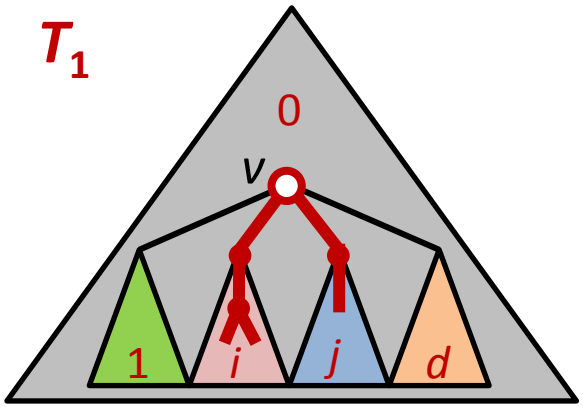
Reduce recoloring cost in T_2 to O(n·log^2 n)



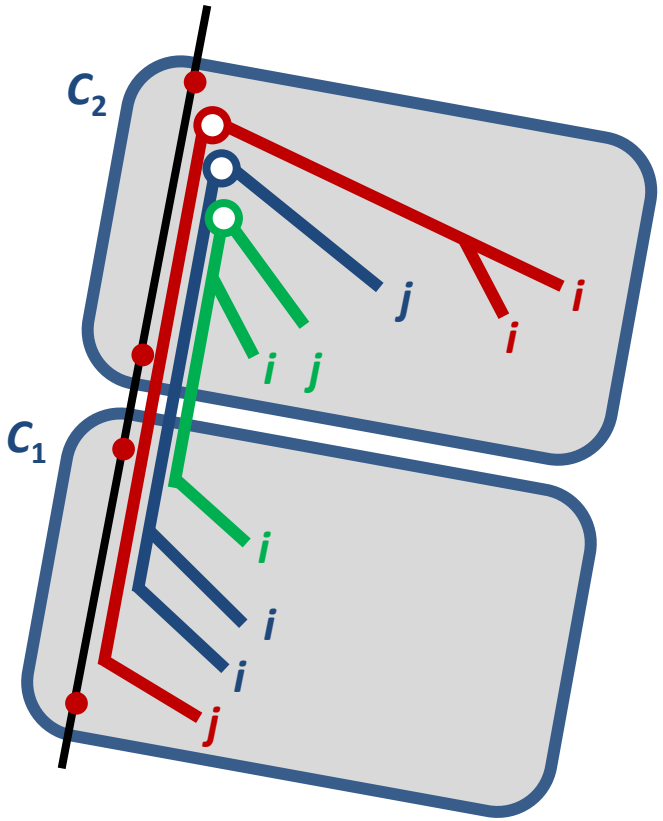
Reduce recoloring cost in T_2 from O(n·log^2 n) *to* O(n·log n)

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

Counting Agreeing Triplets (II)



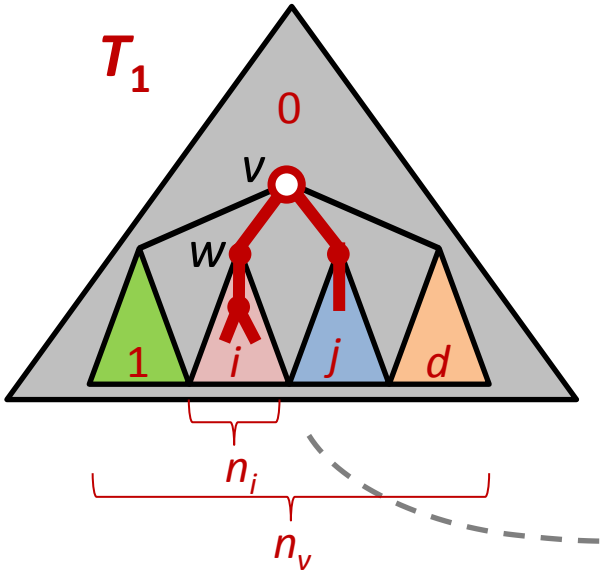
node in $H(T_2)$ = component composition in T_2



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

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

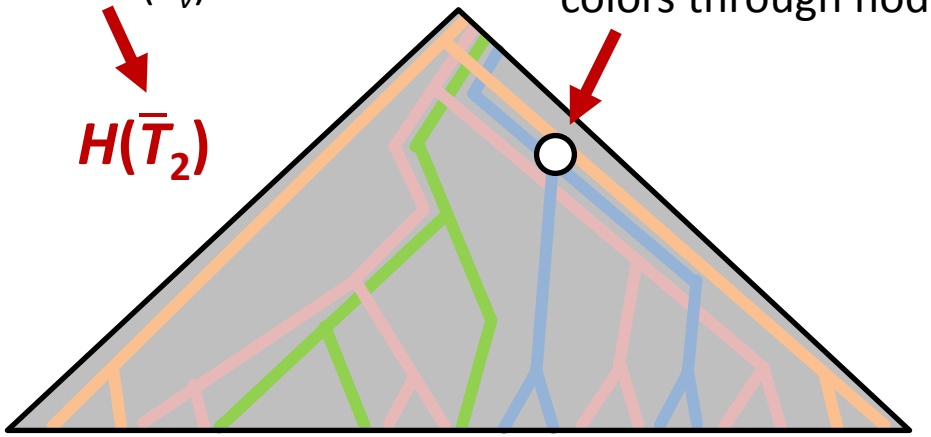
From O(n·log² *n*) to O(*n*·log *n*)



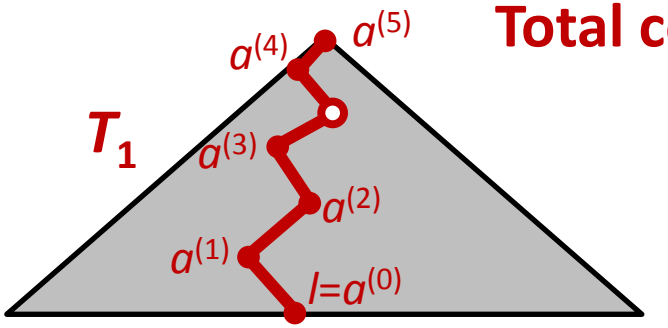
Compressed version of T_2 of size $O(n_v)$

$H(\bar{T}_2)$

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



Colored path lengths $\sum_{2\le i\le d} \log \binom{|\bar{T}_2|}{n_i} = \sum_{2\le i\le d} n_i \cdot \log \frac{n^v}{n_i}$



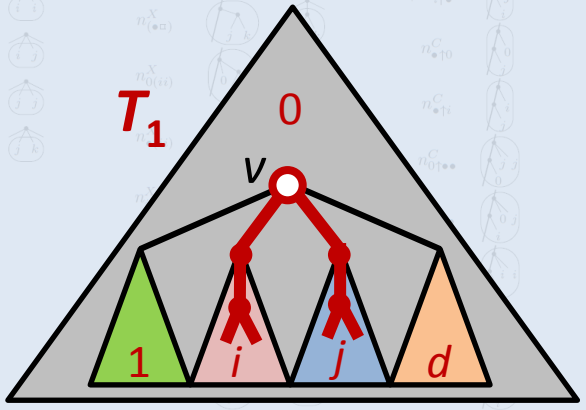
Total cost for updating counters

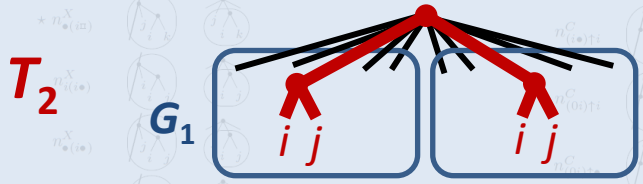
$$\sum_{\text{leaf } l \in T_1} \sum_{\substack{\text{ancestor } a^{(j)} \\ \text{not heavy child}}} \log \frac{n^{a^{(j+1)}}}{n^{a^{(j)}}} = n \cdot \log n$$

Counting Quartets...

- Root T_1 and T_2 arbitrary
- Keep up to $15+38d$ different counters per node in $H(T_2)$...

Bottleneck in computing disagreeing resolved-resolved quartets



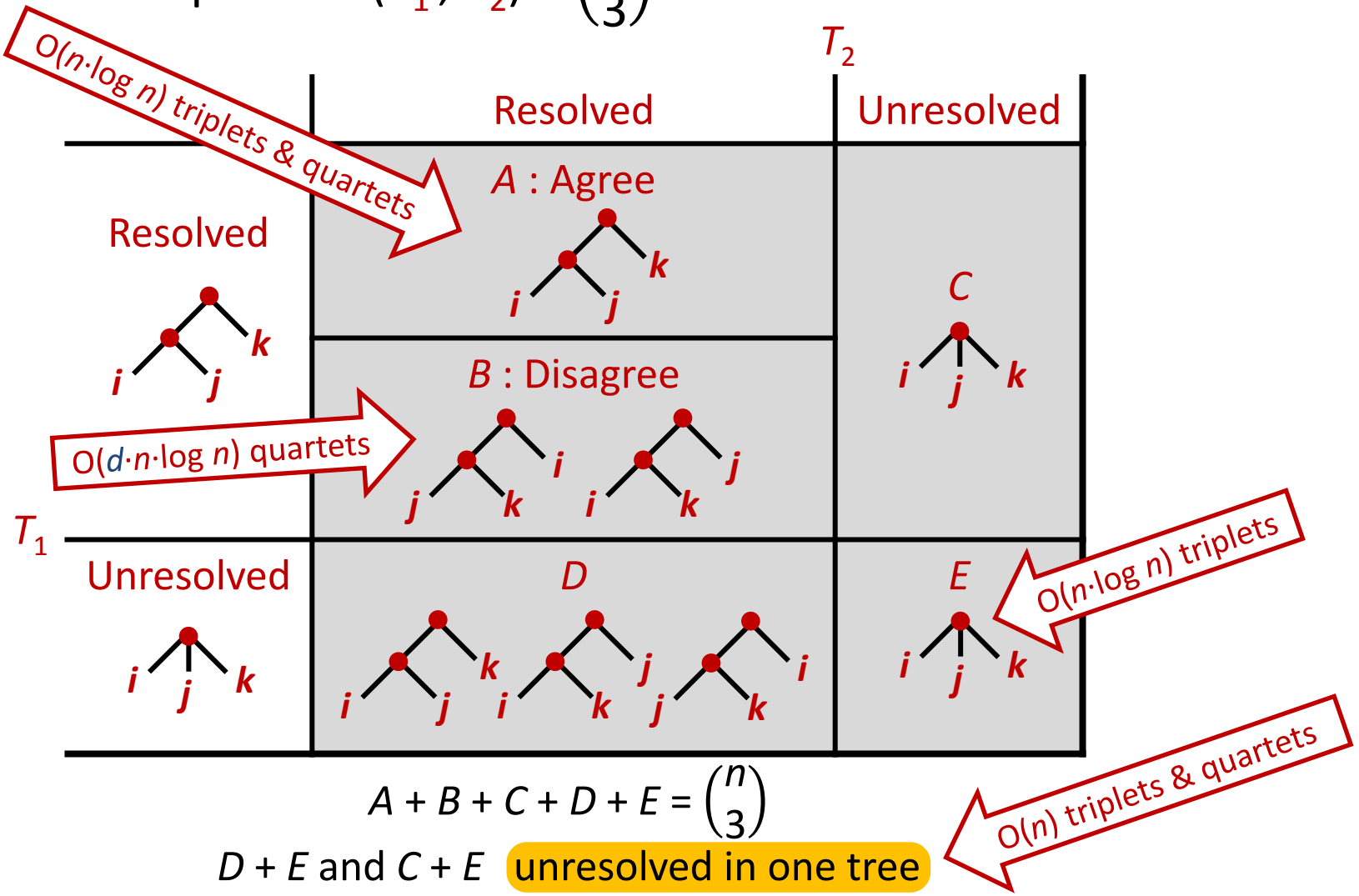


$$\sum_{1 \le i < d} \; \sum_{i < j \le d} n_{(ij)}{}^{G_1} \cdot n_{(ij)}{}^{G_2}$$

double-sum $\Rightarrow$ factor d time

Distance Computation

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

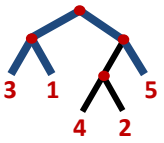
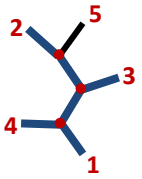
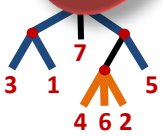
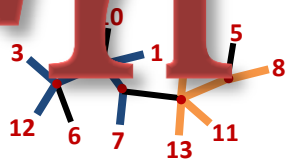


$A + B + C + D + E = \binom{n}{3}$

$D + E$ and $C + E$ unresolved in one tree

Sufficient to compute A and E or A and B

Summary

	Rooted Triplet distance	Unrooted Quartet distance
Bina		
	O(*n*³) CPQ 1996	O(*n*³) D 1985
o(*n*·log *n*) ?	O(*n*·log *n*) [SODA 2013]	O(*n*²) BTK 2000
		O(*n*·log² *n*) BFP 2001
		O(*n*·log *n*) BFP 2003
Degrees ≤ *d*		
	O(*n*²) BDF 2011	O(*d*⁹·*n*·log *n*) SPMBF 2007
	O(*n*·log *n*) [SODA 2013]	O(*n*^2.688) NKMP 2011
		O(*d*·*n*·log *n*) [SODA 2013]

The End

d = maximal degree of any node in T_1 and T_2

O(*n*·log *n*) ?