

Improved Approximation Algorithms for Agreement Forests on k Phylogeny Trees

Asish Mukhopadhyay
University of Windsor

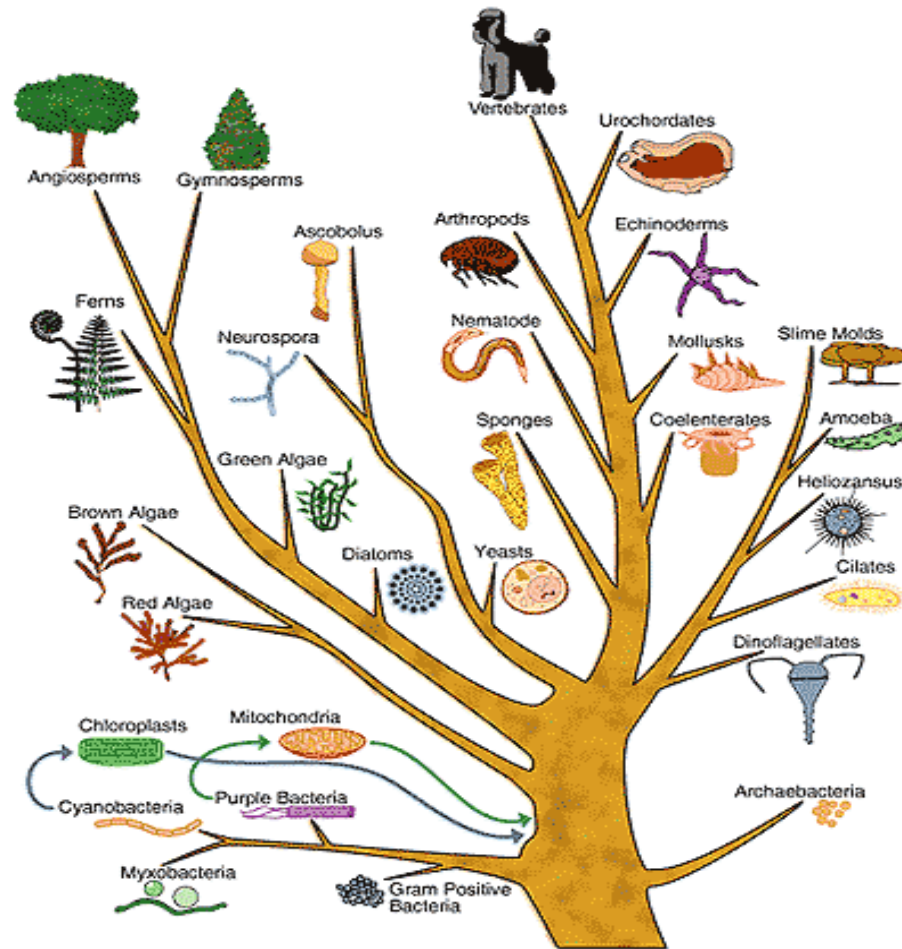
Outline of Talk

- Phylogenetic Trees – What ? & Why ?
- Phylogenetic trees and rSPR-distance
- Measuring their agreement
 - Maximum Agreement Forest (MAFs)
- MAFs and rSPR-distance
- Phylogenetic networks – What ? Why ?
- Phylogenetic networks and hybridization number
- Maximum Acyclic Agreement Forest (MAAFs)
- MAAFs and hybridization number
- Our contributions
- Carrying on.....

Phylogeny Trees – What ? Why ?

- Used for modelling *evolutionary relationships* among a set of biological species
- Also called *evolutionary trees*

Phylogeny Trees – What ? Why ?



Tree-like
evolution of
present day
species

A Phylogeny Tree, Darwin 1837 (Source: Wikipedia)

Phylogeny trees: What ? Why ?

- The evolutionary history of a current set of species can be modelled by *more than one* phylogeny tree
- Gives rise to several interesting problems
 - Compute distance metrics defined on the space of phylogeny trees
 - Find hybrid networks with the fewest hybridization events that explain 2 or more phylogeny trees in terms of non-tree-like or *reticulate* evolution

rSPR Operations on Phylogeny trees

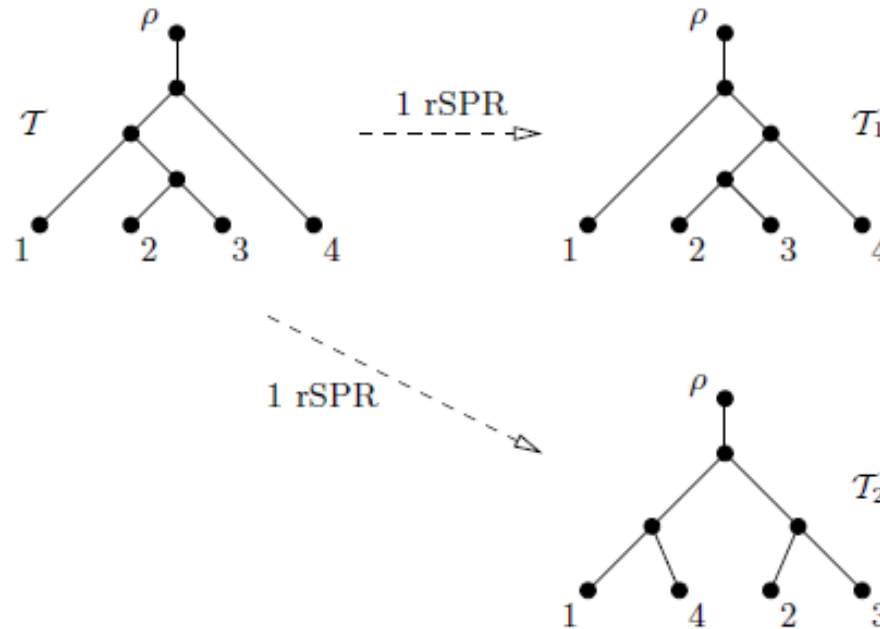


FIGURE 1. Each of T_1 and T_2 are obtained from T by a single rSPR operation.

Figure Credit: A 3-approximation algorithm for the subtree distance between Phylogenies, M. Bordewich, C. McCartin and C. Semple, J Disc. Algos, 2008, 458-471

Phylogeny trees and *rSPR-distance*

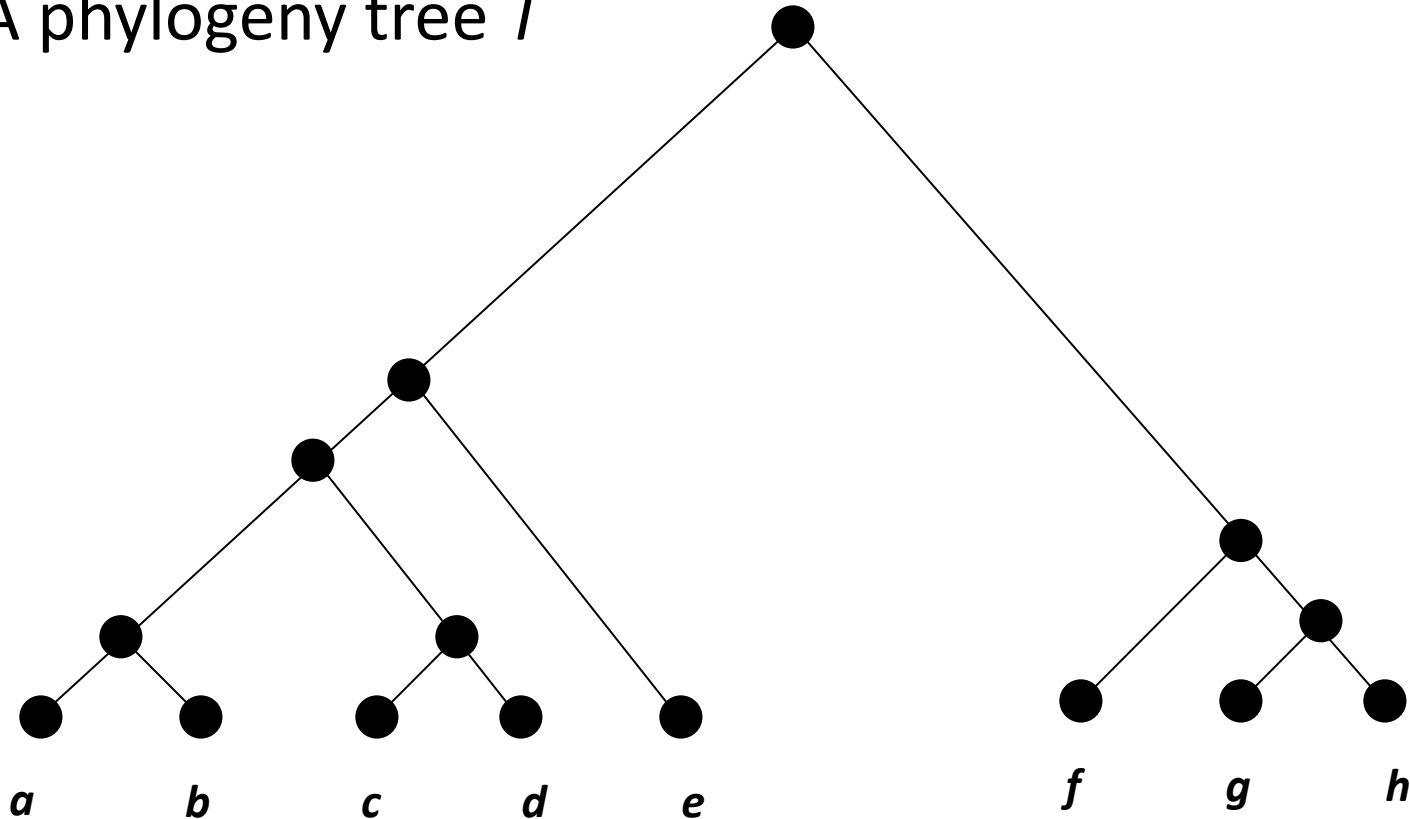
- The *rSPR-distance* is a distance metric defined on the space of phylogeny trees
- Measures the # of *rSPR* operations needed to morph one phylogeny tree into another
- Computing *rSPR-distance* ?

Measuring their Agreement

- With the help of Agreement Forests
- Roughly speaking , an Agreement Forest is..
 - A bunch of common sub-trees of 2 phylogeny trees
 - Greater the similarity, fewer the common sub-trees
 - A Maximum Agreement Forest (MAF) has the fewest common sub-trees

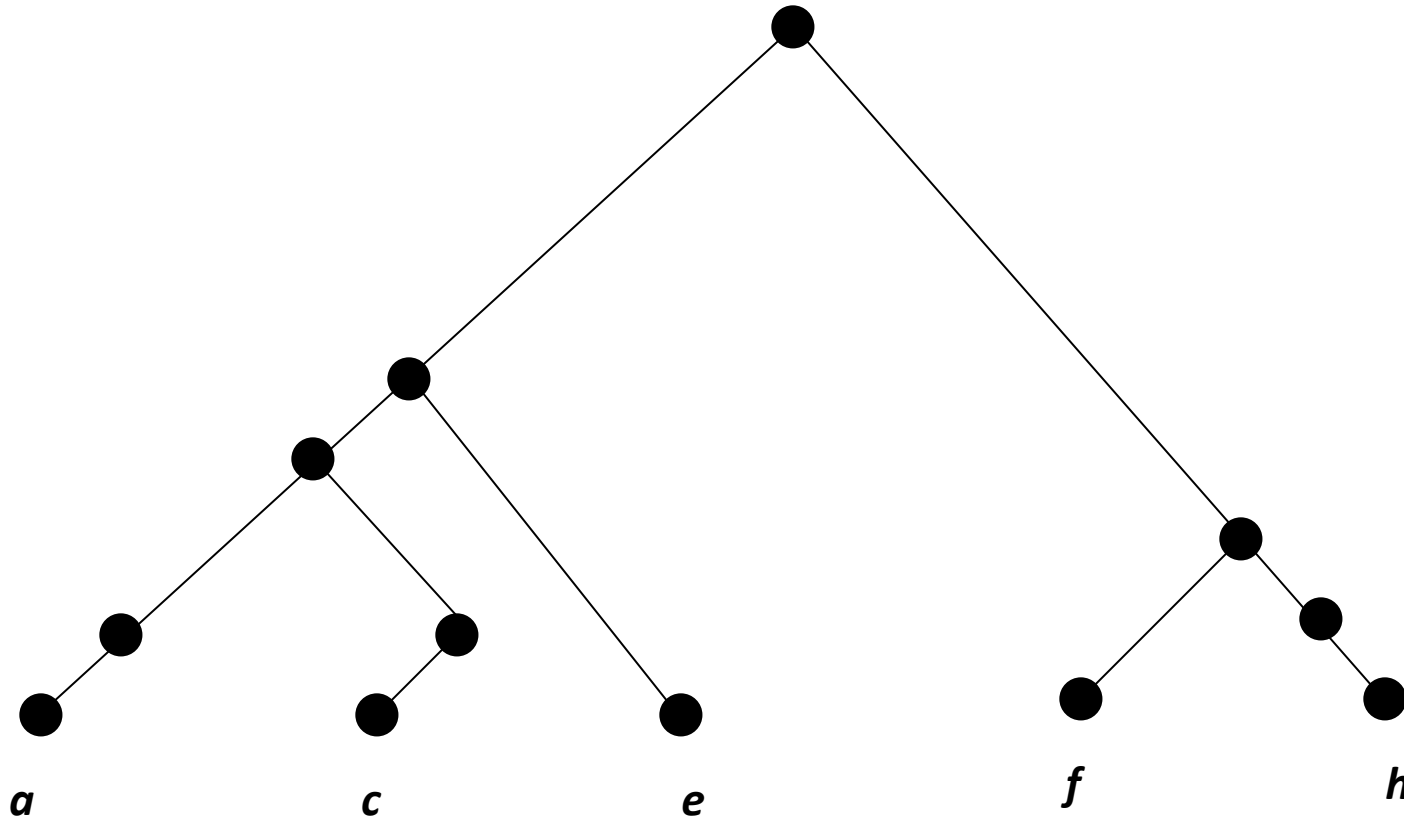
Measuring their Agreement

A phylogeny tree T



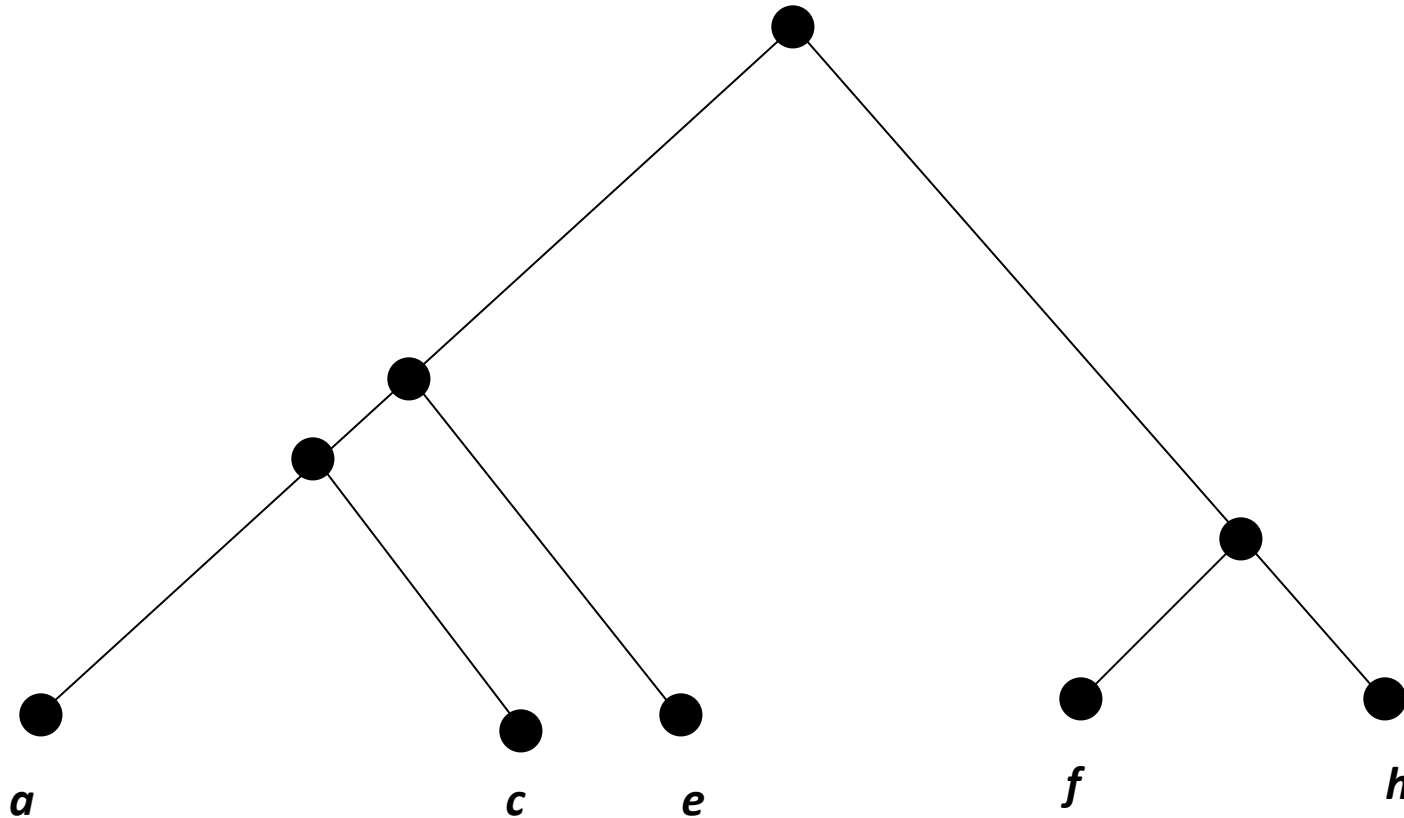
Leaf set $L(T) = X = \{a, b, c, d, e, f, g, h\}$

Measuring their Agreement



$X' = \{a, c, e, f, h\}$ and $T(X')$

Measuring their Agreement

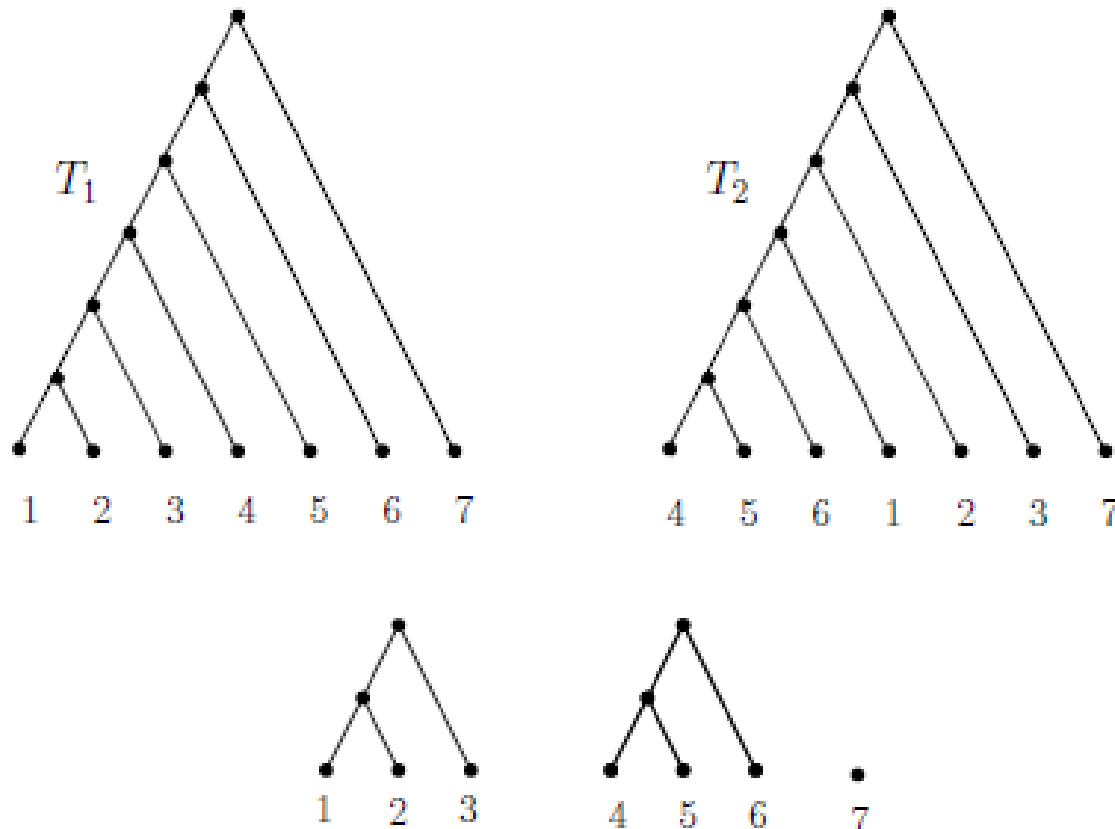


$$X' = \{a, c, e, f, h\} \text{ and } T|X'$$

Measuring their Agreement

- Formally, t_1, t_2, \dots, t_k is an *agreement forest* of T_1 and T_2 if
 - $\bigcup L(t_i) = L(T)$
 - For all $i \in \{1, 2, \dots, k\}$, $t_i \approx T_1 | L(t_i) \approx T_2 | L(t_i)$
 - The graphs $T_1(L(t_i))$, $i=1, 2, \dots, k$ are vertex-disjoint
 - As also the graphs $T_2(L(t_i))$, $i=1, 2, \dots, k$

Measuring their Agreement



Measuring the agreement of T_1 and T_2

MAFs and rSPR-distance

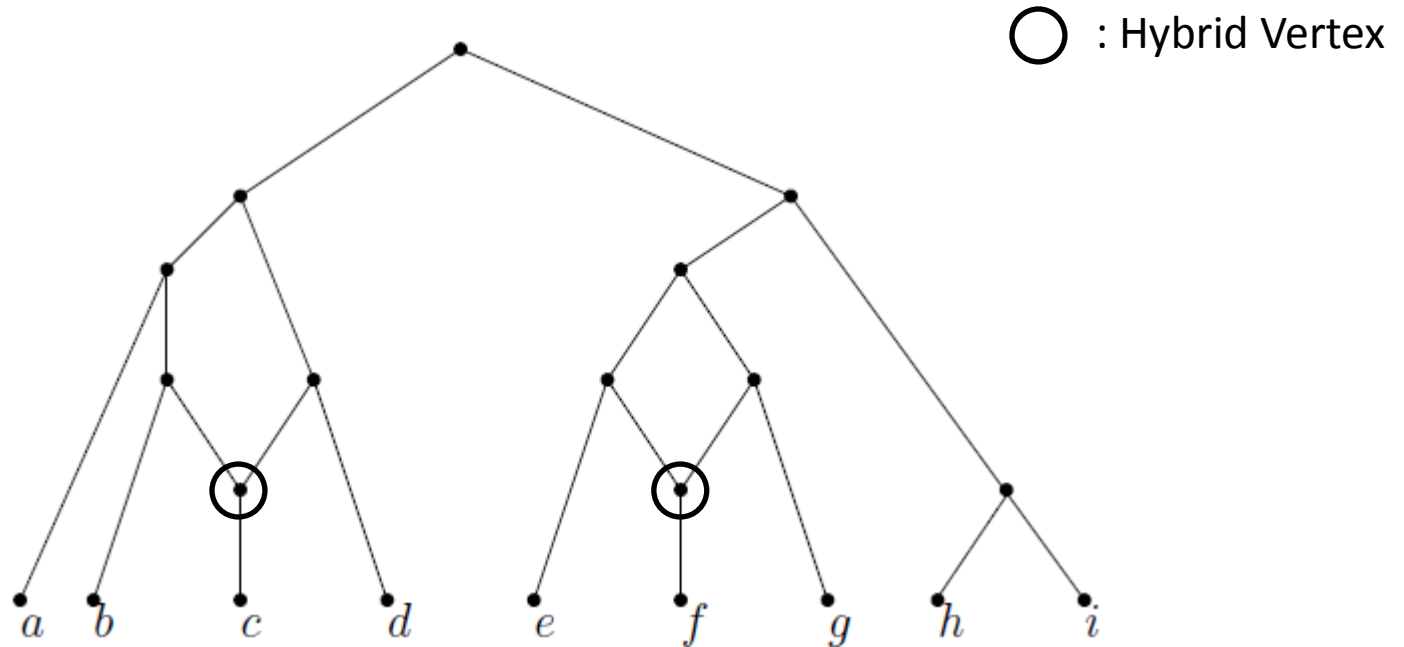
Theorem:

Minimum rSPR-distance = Size of a MAF - 1

Phylogeny networks – What ? Why ?

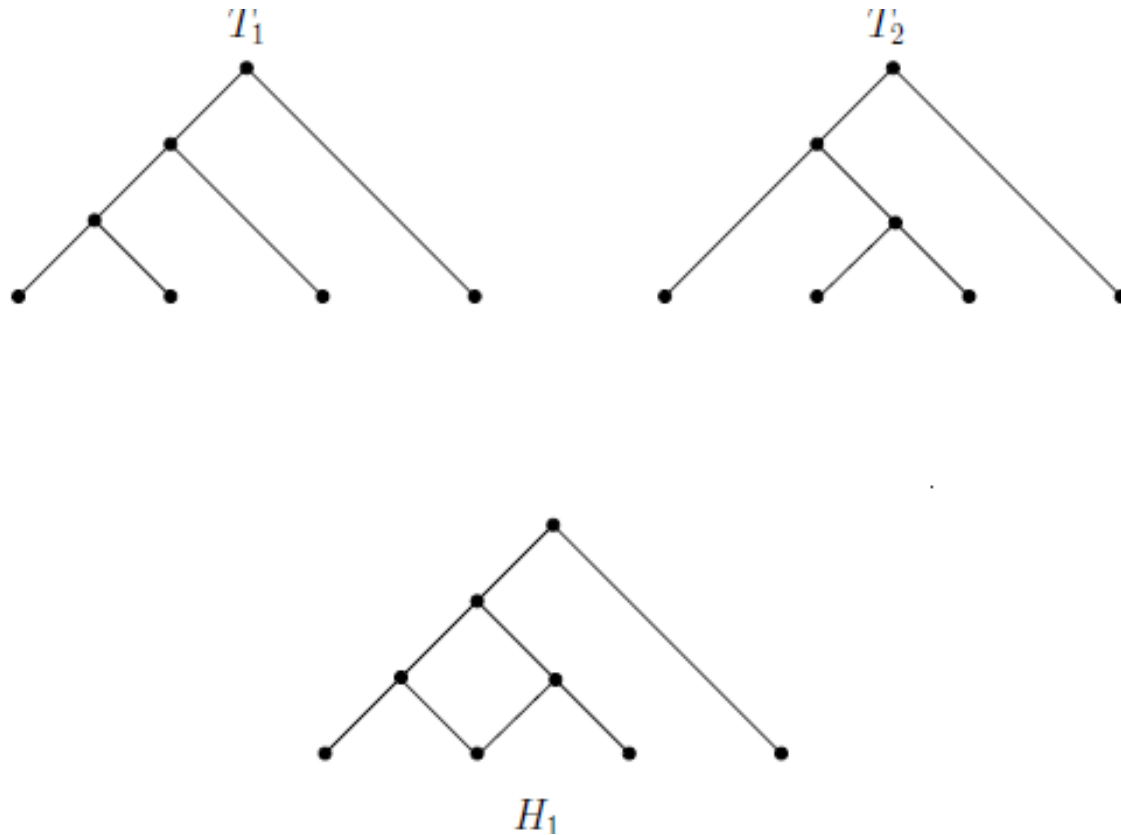
- Also called Hybrid Networks
- Reticulate Evolution
 - Non-tree like evolution, covering
 - Hybridization
 - Lateral Gene Transfer
 - Recombination
 - Modelled by
 - Phylogeny Networks

Phylogeny Networks – What ? Why ?



A Phylogeny Network is a DAG

Phylogenetic Networks – What ? Why?



Phylogeny network H_1 represents the phylogeny trees T_1 and T_2

Phylogeny Networks – What ? Why?

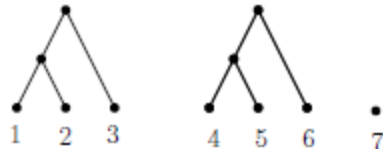
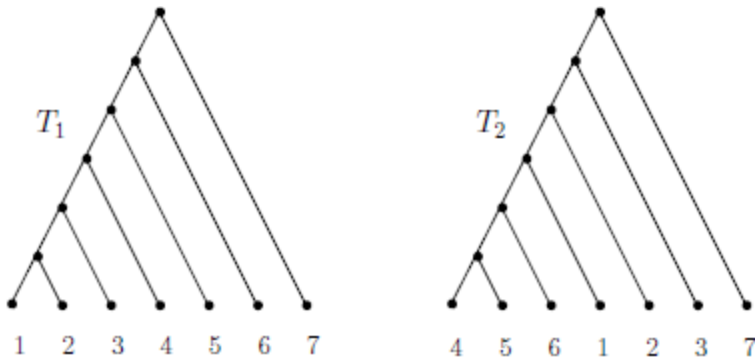
- Hybridization Number:
 - Minimum number of hybrid vertices needed to represent 2 or more phylogeny trees
 - **Notation:** $h(T_1, T_2)$ for 2 phylogeny trees
- Our focus:
 - Computation of Hybridization Number

Maximum Acyclic Agreement Forest

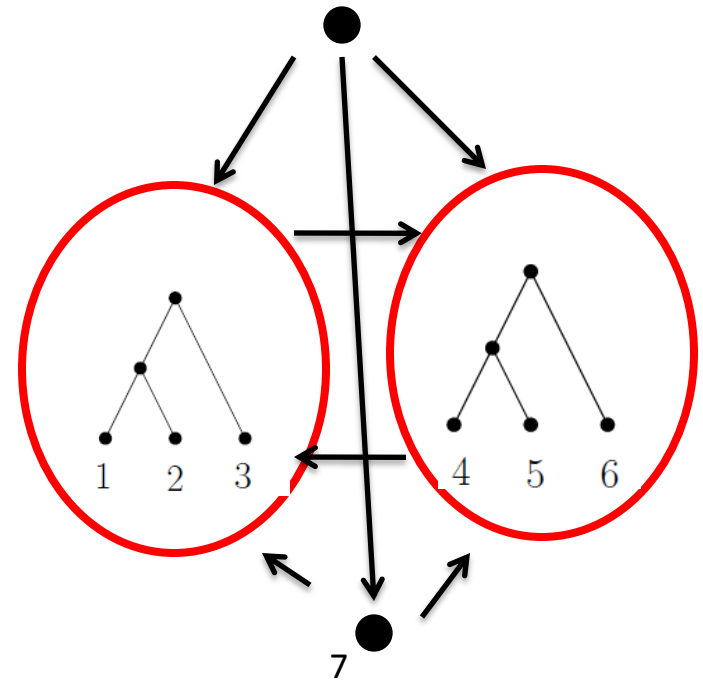
- A MAF is *acyclic* if there are no 2 trees t_i and t_j in the forest such that
 - the root of t_i is an ancestor of t_j in one phylogenetic tree and
 - t_j is an ancestor of t_i in the other

Cycles in an agreement forest

Phylogeny trees

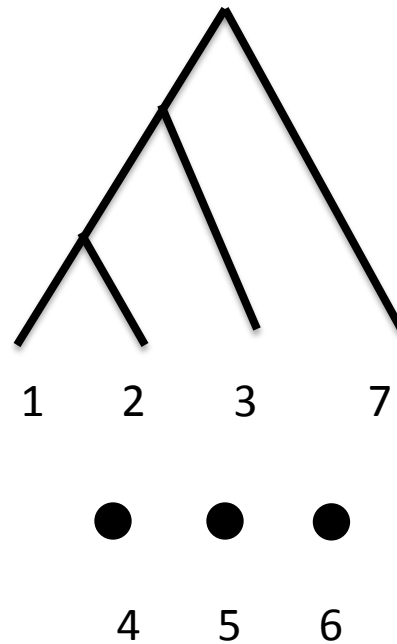
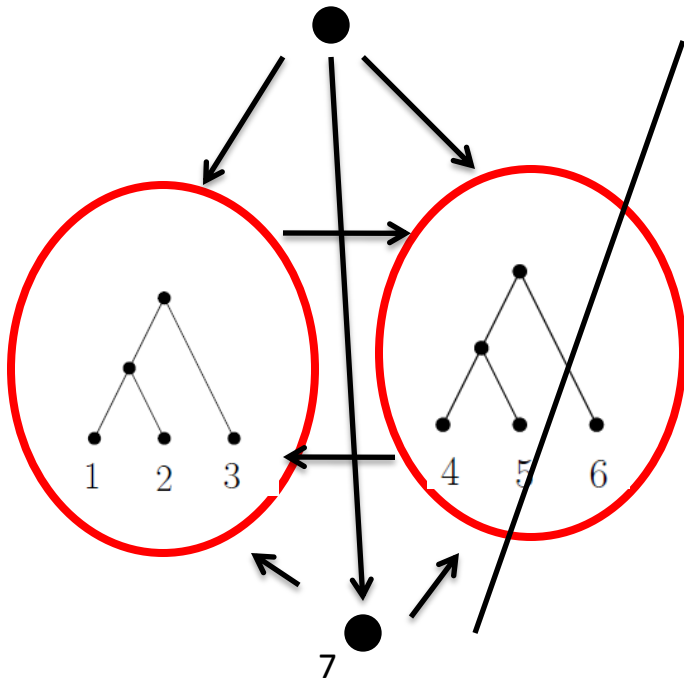


Agreement Forest



Cycle Graph

Maximum Acyclic Agreement Forest



One of the trees in the cycle can't be realized

MAAF

MAAF and Hybridization Number

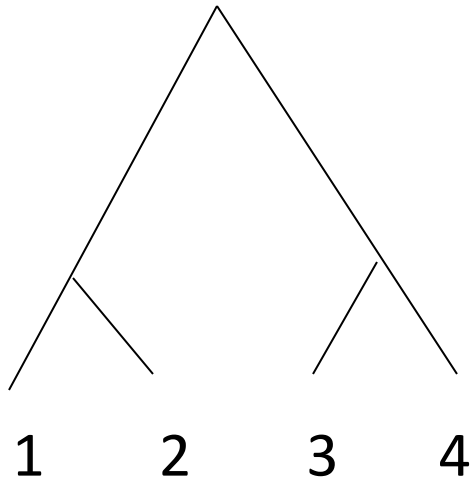
Theorem:

$$h(T_1, T_2) = \text{size of a MAAF} - 1$$

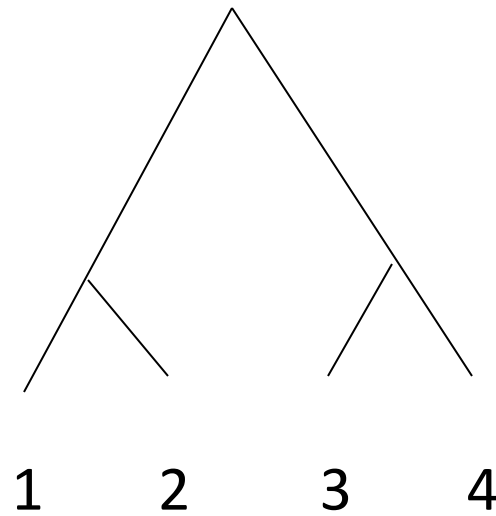
Computing a MAF

- Computing a MAF of 2 trees T_1 and T_2 is NP-hard (Hein et al. ,1996)
- Hence the interest in computing a MAF approximately

Computing a MAF approximately



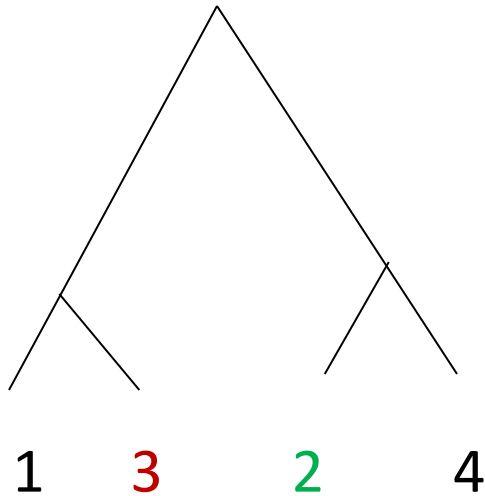
T_1



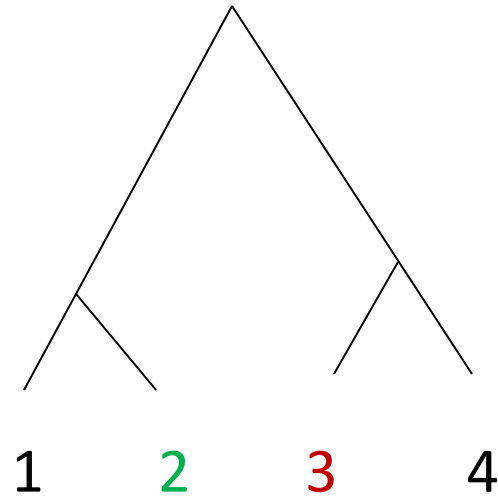
T_2

Two trees with *identical* topology

Computing a MAF approximately



T_1



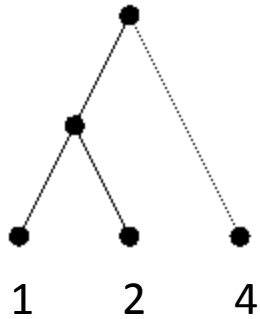
T_2

Two trees with *different* topology

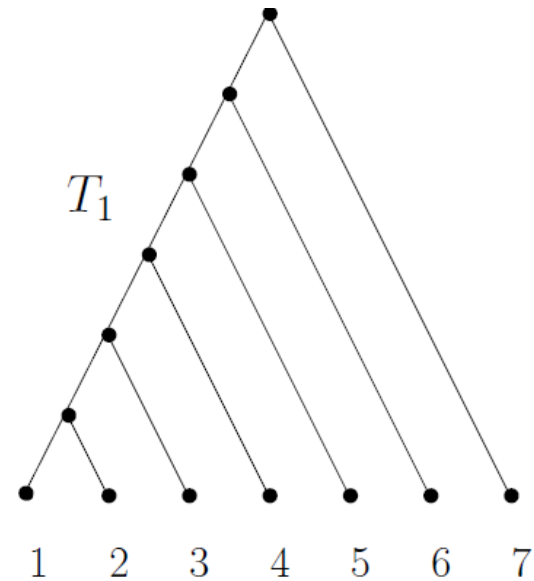
Computing a MAF approximately

- P_{topo}
 - Compute a forest F by eliminating edges of T_1
 - Each tree in F occurs with the same topology in both T_1 and T_2
- Overlaps
 - Eliminate more edges of F to remove overlaps in T_2 of pairs of trees in F

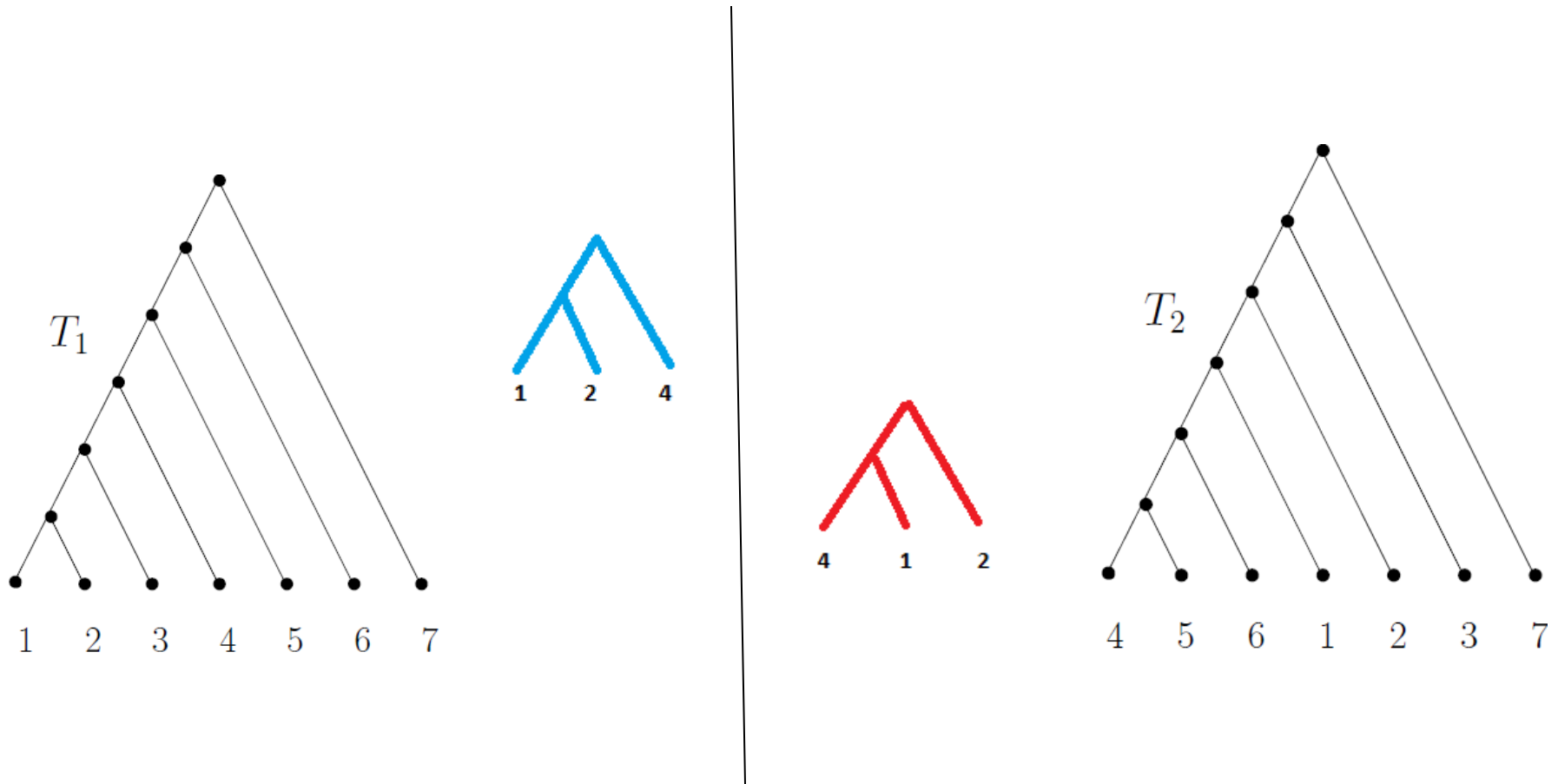
P_{topo} : Triples



is a triple of



P_{topo} : Incompatible Triples (ITs)



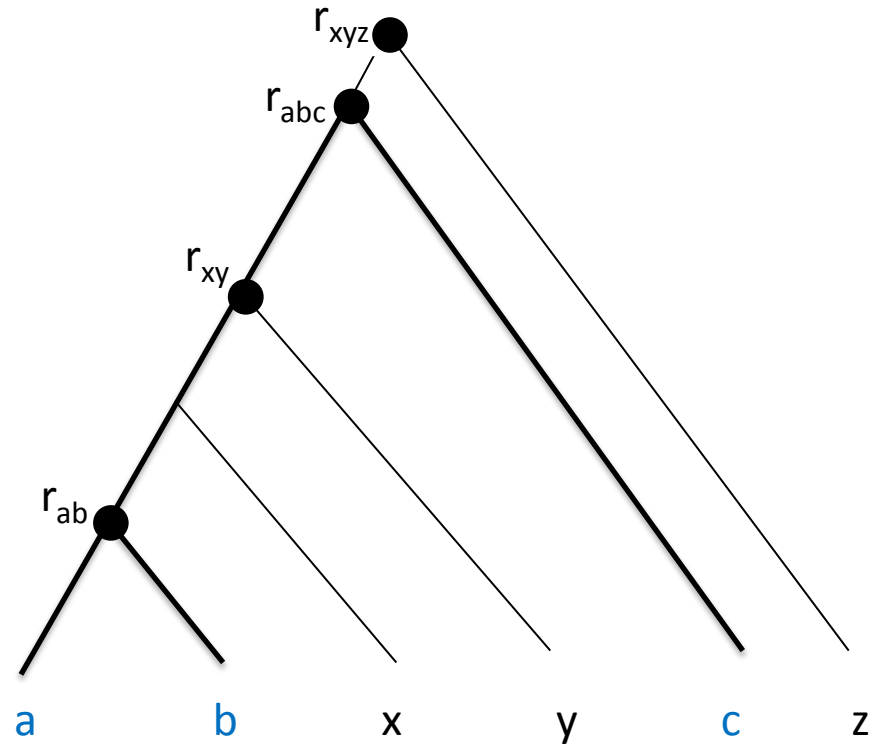
The blue triple 12|4 of T_1 is incompatible with the red triple 41|2 of T_2

P_{topo} : Partial Order on ITs

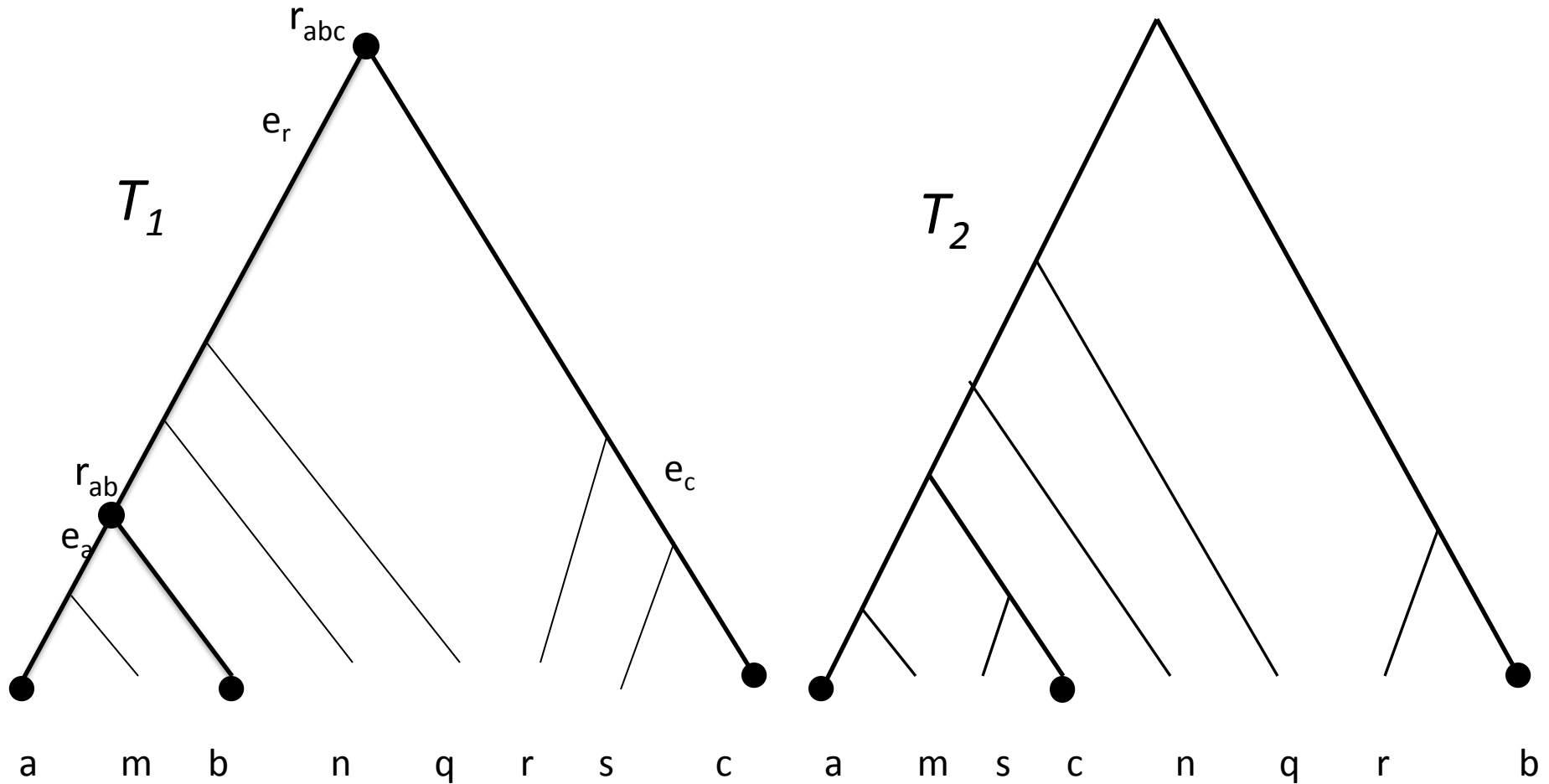
$ab|c < xy|z$ if

i) either r_{abc} is a descendant of r_{xyz}

ii) or if $r_{abc} = r_{xyz}$, r_{ab} is a descendant of r_{xy}

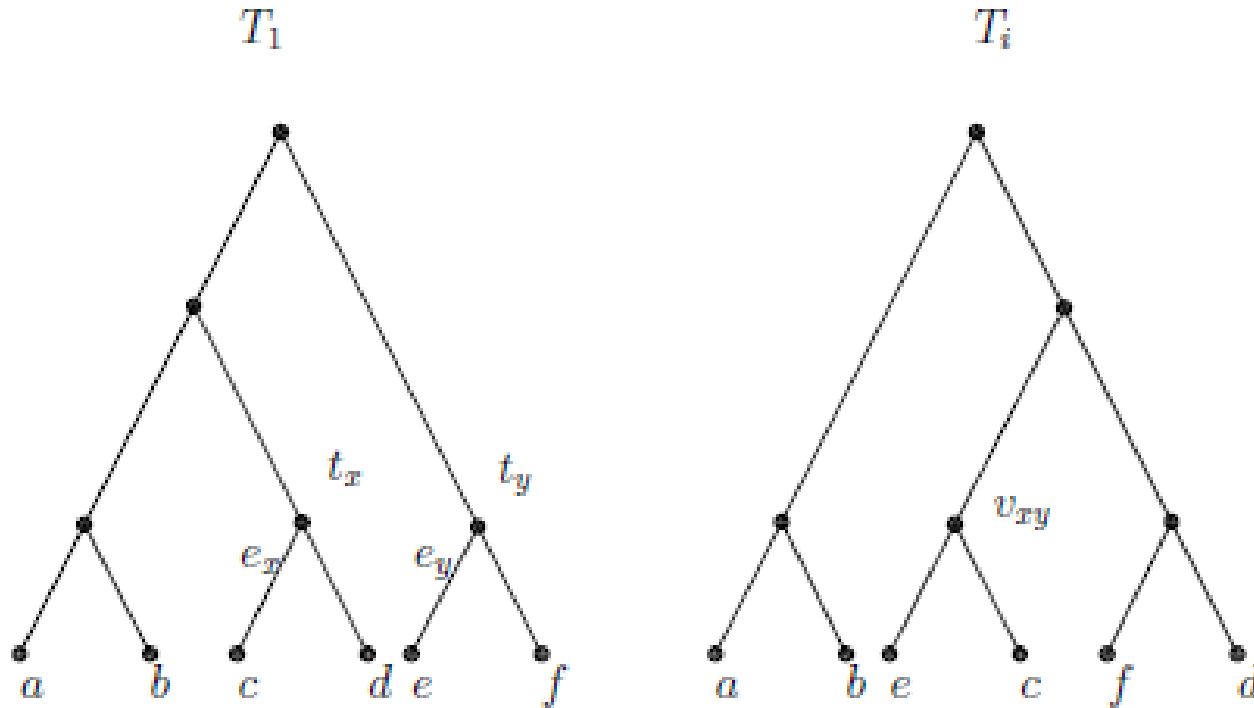


P_{topo} : Remove incompatibility



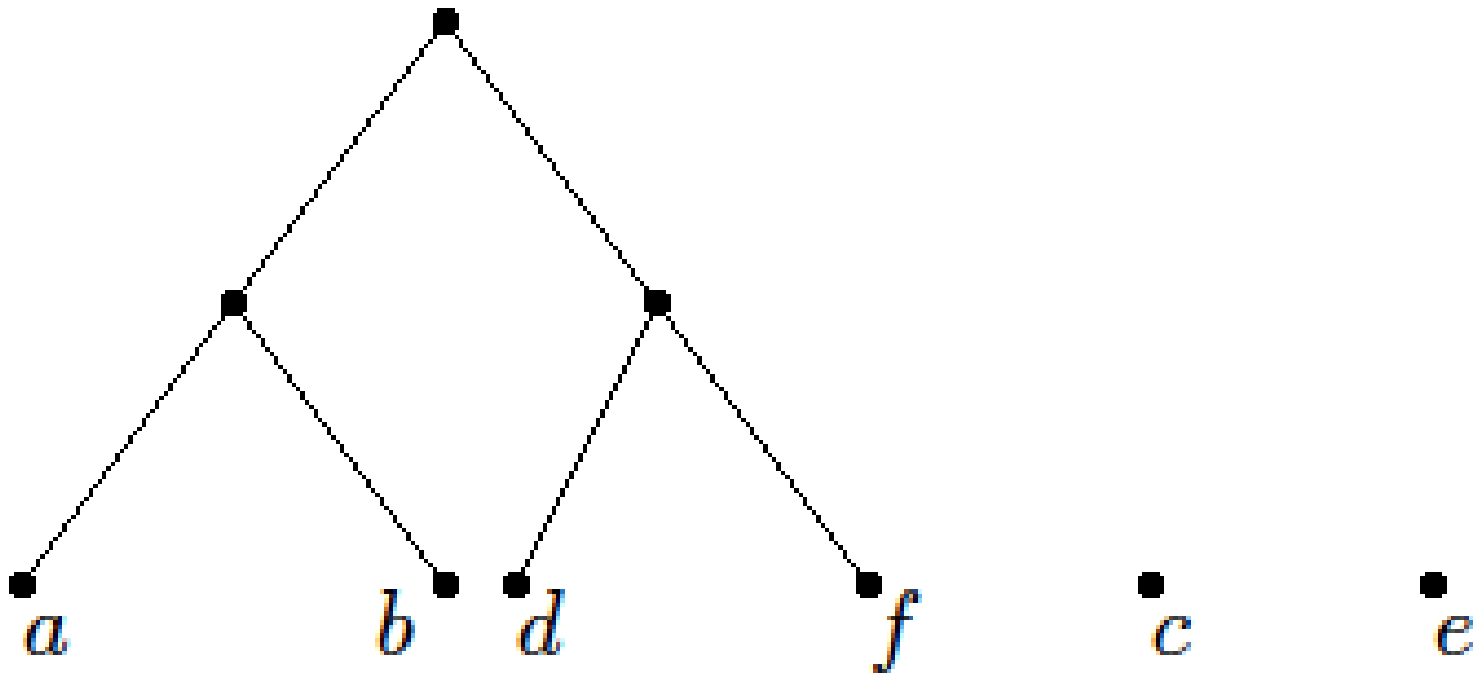
$ab|c$ and $ac|b$ are incompatible triples; edges e_a , e_r and e_c are deleted from T_1

Component Overlap



Components t_x and t_y of $F(=T_1)$ overlap in T_i , for some $i \geq 2$

Resolving the Overlap

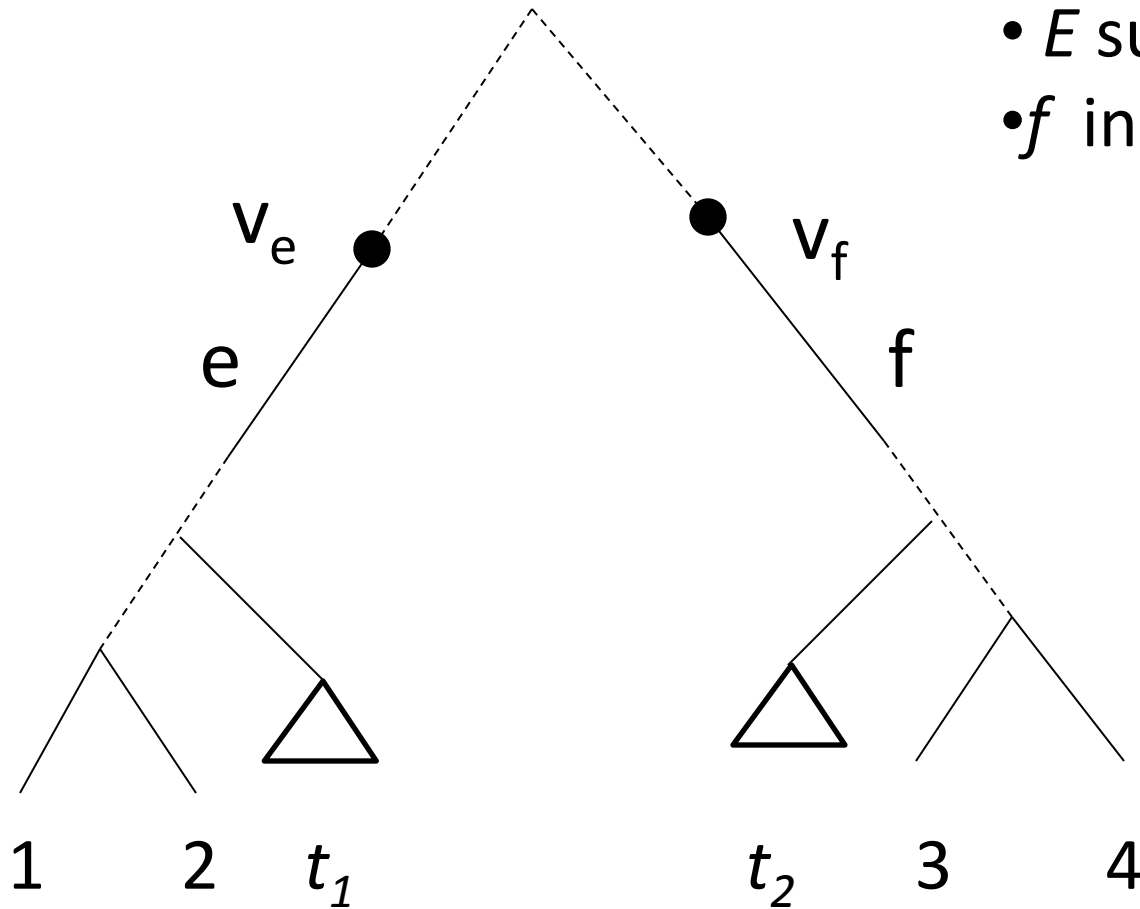


Edges e_x and e_y deleted to resolve component overlap

Edge Deletion : Underlying Theory?

- Let
 - F be a forest obtained from a phylogeny tree T_1 by *deleting* some of its edges and *contracting* degree 2 vertices
 - E be a subset of edges of F

Conditions of Shifting Lemma

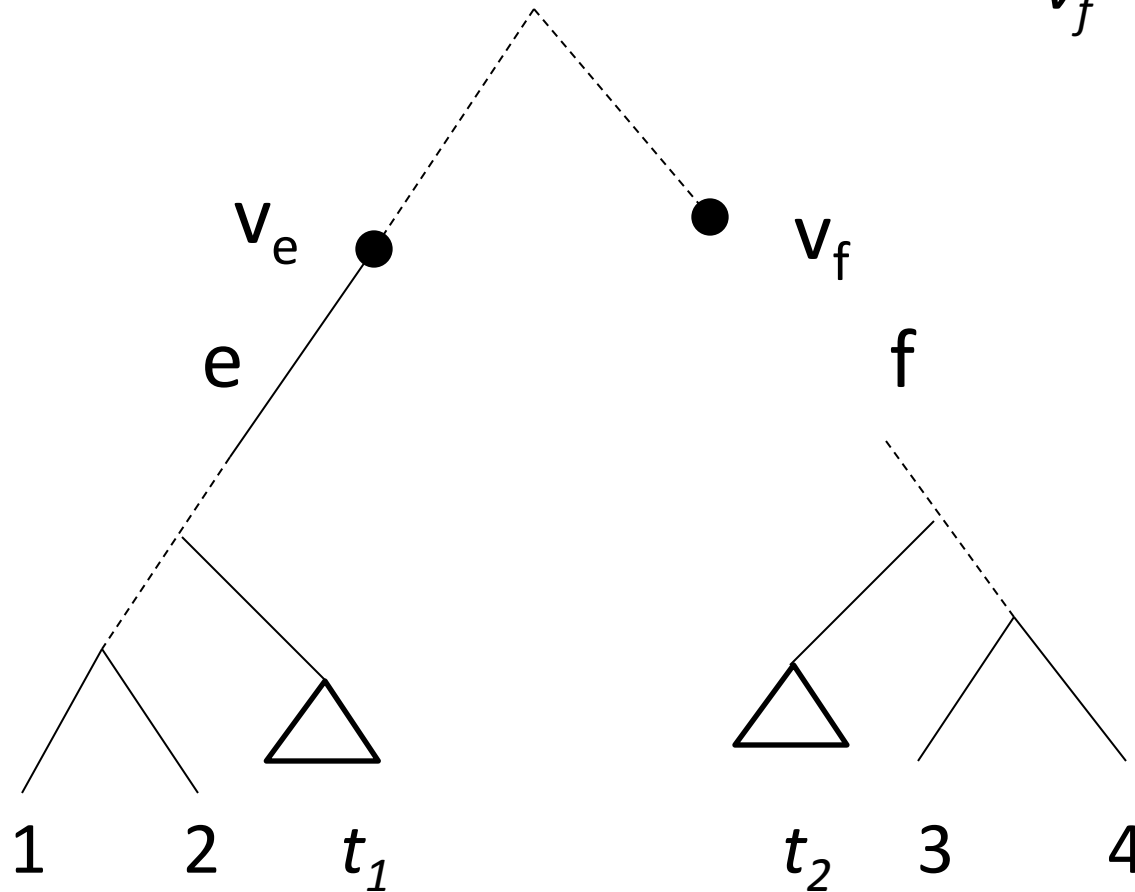


- E subset of F
- f in E , e not in E

A tree t in the forest F

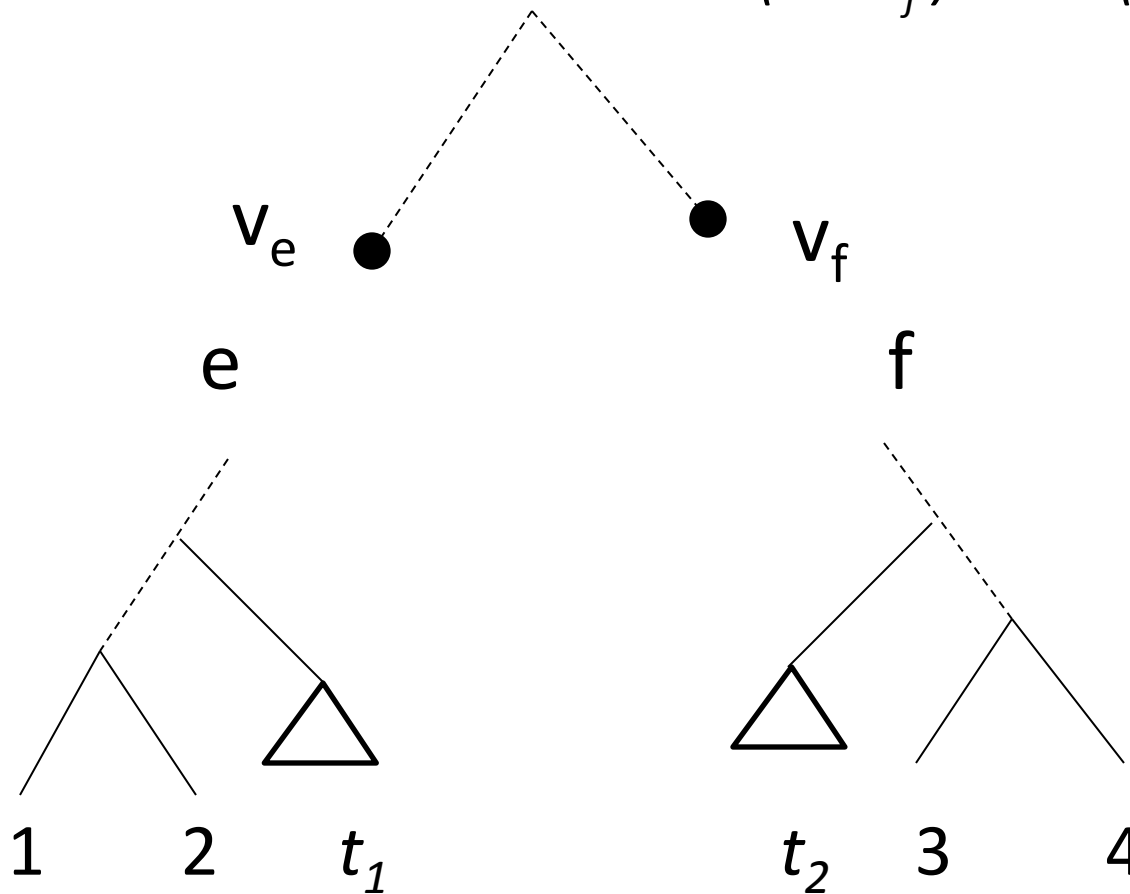
Conditions of Shifting Lemma

• $v_f \sim v_e$ in $F - E$



Conditions of Shifting Lemma

- $\neg (a \sim v_f)$ in $F - (E+e)$ for all $a \in X$



Conclusion of Shifting lemma

- The forests $F - E$ and $F - (E - f + e)$ are isomorphic
- This lemma underlies the next two theorems that explains the removal of
 - edges e_a and e_r from F , corresponding to a minimum incompatible triple
 - edges e_x and e_y when two components in the current forest overlap in T_2

Two theorems

- Theorem 1:
 - Let E be the minimum set of edges such that $F-E$ yields an agreement forest of T_1 and T_2 , where $F = T_1$ initially.
 - If $ab|c$ is an incompatible triple of F with respect to T_2 , then there exists an edge f in E such that $F - (E - f + \{e_a, e_r\})$ is *isomorphic* to a sub-forest of $F-E$

Summary of Theorem 1

- In lieu of removing the *unknown* edge f from F we can remove the *known* edges e_a and e_r with practically the same effect

Two theorems

- Theorem 2:
 - Assume F has no triple incompatible with T_2 .
 - If T_s and T_t are two components of F that overlap in T_2 , then there exists an edge f in E such that for some i in $\{s,t\}$, $F - (E - f + e_i)$ is isomorphic to $F - E$.

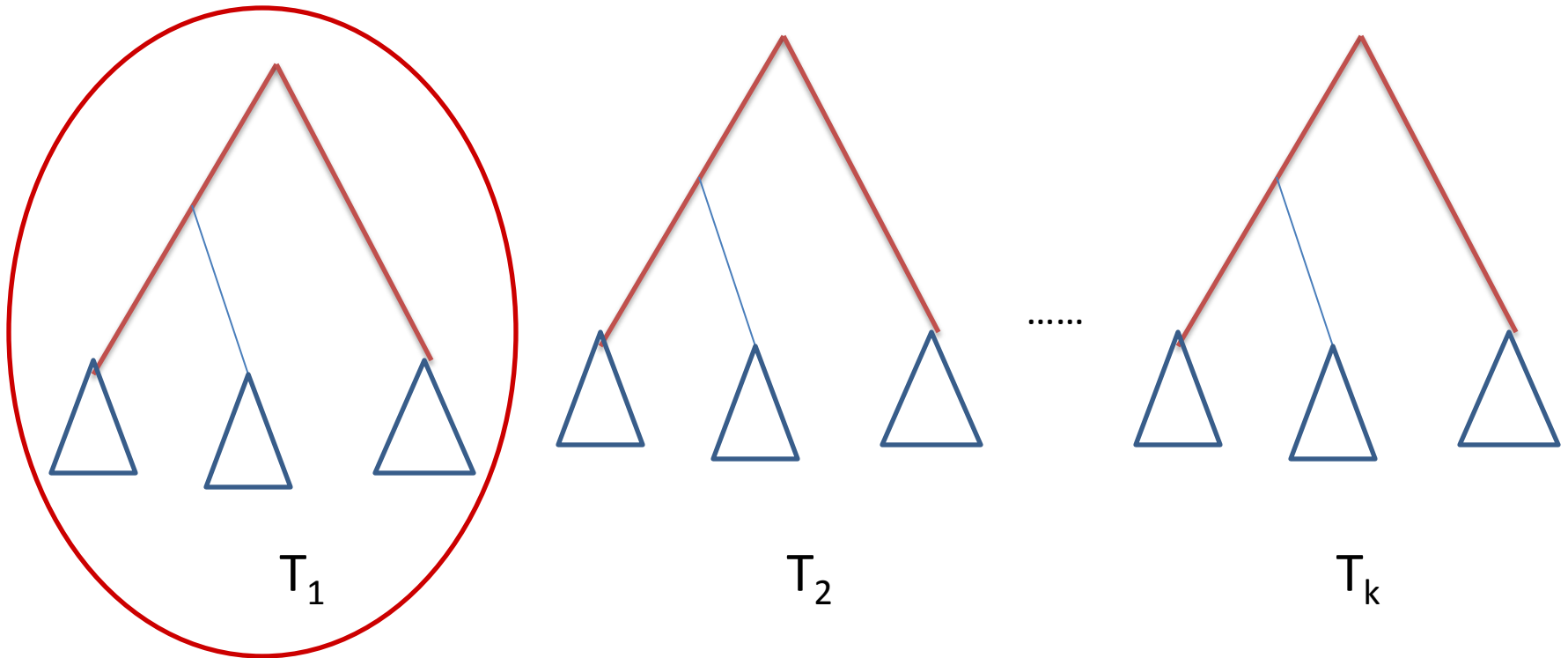
Summary of Theorem 2

- In lieu of removing the *unknown* edge f from F we can remove the *known* edges e_s and e_t with practically the same effect

Approximation Algorithm for computing MAF on k trees

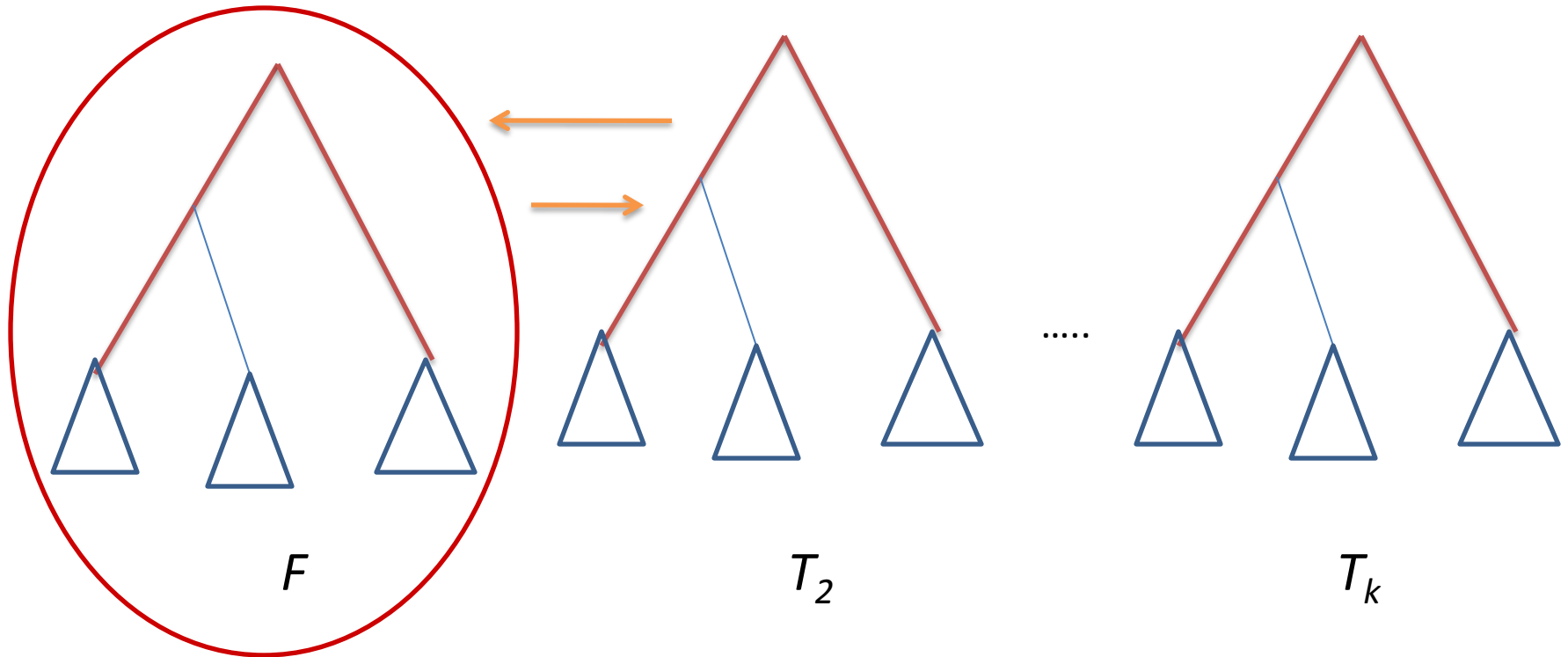
- The last two theorems form the basis for a 2-approximation algorithm

Algorithm, Pictorially



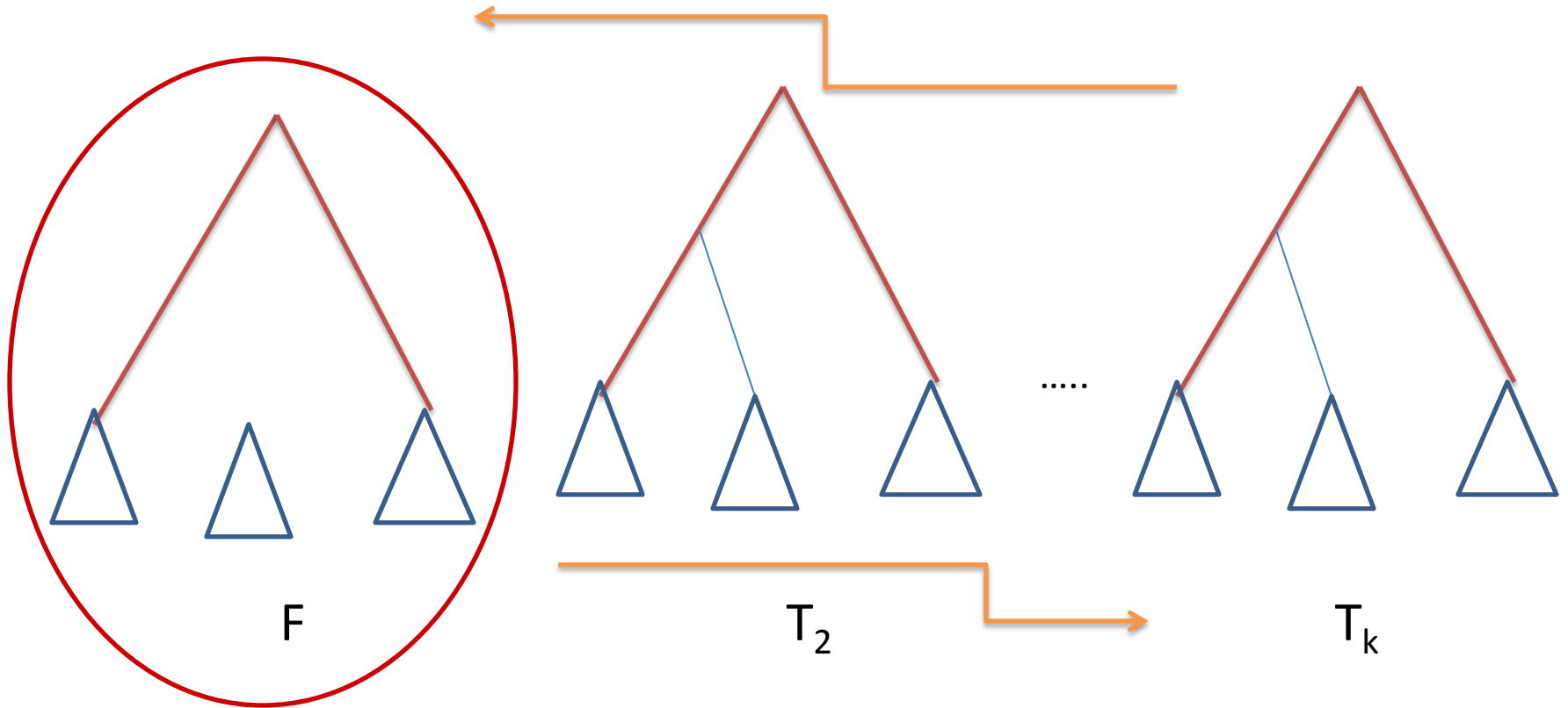
We fix one tree T_1 and make a forest F out of T_1

Algorithm, Pictorially



At a time we check the incompatibility and overlapping components between two trees of F and a T_i ($i \neq 1$) and cut the right edges from F

Algorithm, Pictorially



More Formally.....

Algorithm *MAF-Approx*(T_1, T_2, \dots, T_k)

1. $F \leftarrow T_1$;
2. for $i = 2$ to k
do
 - 2.1. while there exists an incompatible triple in F with respect to T_i
do
 - 2.1.1. consider the minimal incompatible triple $ab|c$ in F with respect to T_i
 - 2.1.2. $E \leftarrow \{e_a, e_r\}$ in $ab|c$
 - 2.1.3. $F \leftarrow F - E$
- 2.1. enddo;
2. enddo;
3. for $i = 2$ to k
do
 - 3.1. while there exists a pair of inseparable components in any T_i ($i \geq 2$) with respect to F
do
 - 3.1.1. consider inseparable components t_x and t_y in T_i with respect to F
 - 3.1.2. $E \leftarrow \{e_x, e_y\}$ in t_x and t_y
 - 3.1.3. $F \leftarrow F - E$
- 3.1. enddo;
3. enddo;
4. return F ;

2-approximation ratio claim

- **Claim:**

- Algorithm MAF-Approximate has approximation ratio 2, improving on a previous ratio 8 algorithm.

- **A notation:**

- Let $e(F, \{T_2, T_3, \dots, T_k\})$ denote the size of a minimum set of edges E of F such that $F - E$ yields *an agreement forest* of F .

Lemma 1

- Let there be k rooted binary phylogenetic trees T_1, T_2, \dots, T_k and let F be a forest of T_1 . If $ab|c$ is a minimal incompatible triple of F with respect to some $T_i, i \geq 2$, then

$$e(F - \{e_c, e_r\}, \{T_2, T_3, \dots, T_k\}) \leq e(F, \{T_2, T_3, \dots, T_k\}) - 1$$

Lemma 1 (contd.)

- If F has *no incompatible triple* with any T_i , but two of its trees t_x and t_y overlap in some T_i then

$$e(F - e_i, \{T_2, T_3, \dots, T_k\}) = e(F, \{T_2, T_3, \dots, T_k\}) - 1$$

– where $i = x$ or y

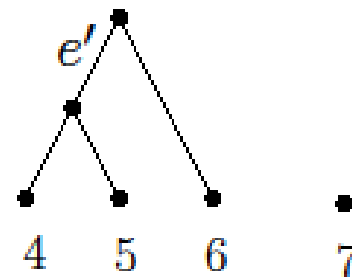
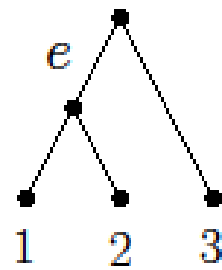
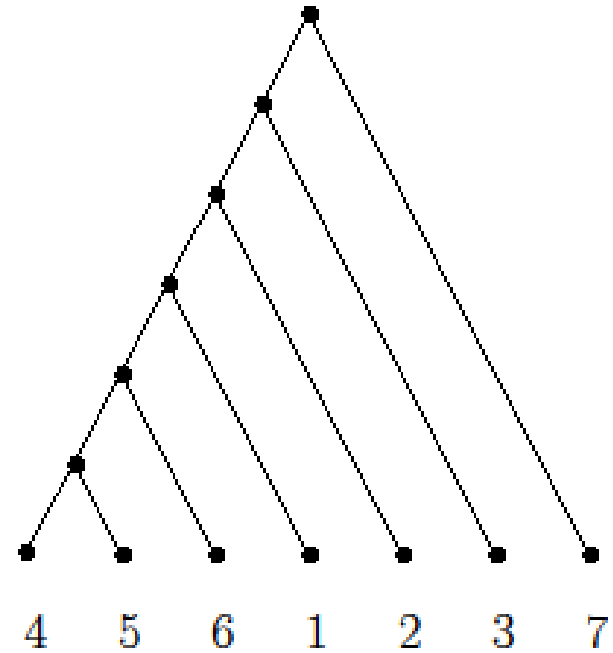
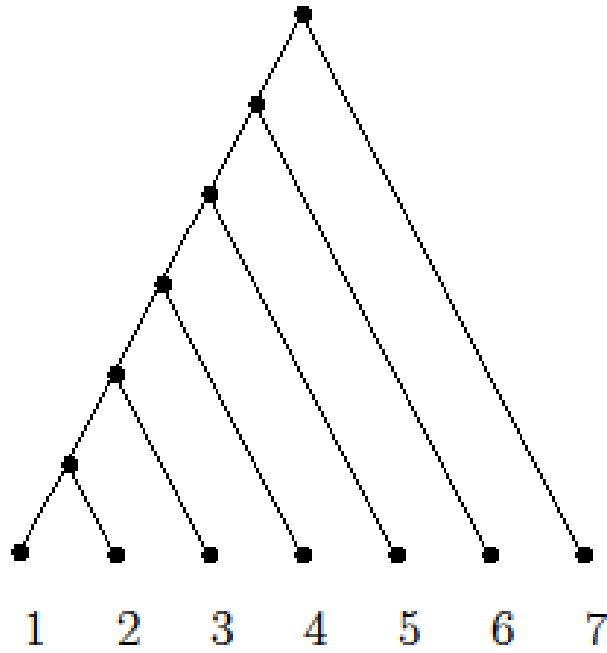
Lemma 1 (contd.)

- We can use this result to show:
 - $\alpha \leq e(T_1, \{T_2, T_3, \dots, T_k\}) \leq 2\alpha$
 - α : total # of iterations of *MAF-approximate*
- Since $e(T_1, T_2, T_3, \dots, T_k) = m(T_1, T_2, T_3, \dots, T_k)$
 - #edges removed = $2\alpha \leq 2 m(T_1, T_2, T_3, \dots, T_k)$,
 - Thus we have a 2-approximation algorithm.

Approximate MAAF on k trees

- Basic idea
 - Start with a 2-approximate MAF produced by our algorithm
 - Eliminate cycles between pairs of trees in this approximate MAF

Example Agreement Forest

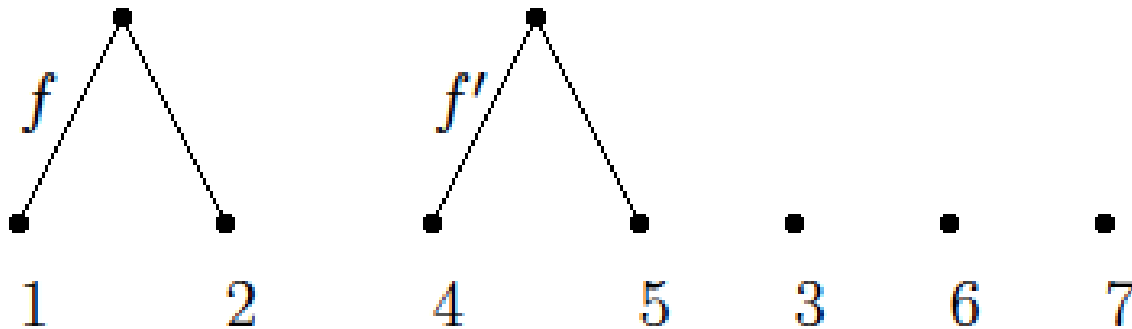


Example Agreement Forest

- Roots of the trees $t(\{1,2,3\})$ and $t(\{4,5,6\})$ are in a *cycle* with respect to their ancestor-descendant relationship

Edge removal

- Remove edges e and e' to eliminate cycle
- This gives us a new forest:



- Roots of the trees $t(\{1,2\})$ and $t(\{4,5\})$ are in a cycle with respect to their ancestor-descendant relationship
- Eliminate edges f and f' to take care of this

Final Agreement Forest

- 1
- 2
- 4
- 5
- 3
- 6
- 7

MAAF- Algorithm

Algorithm *MAAF-Approx(F)*

// $F = \{t_1, t_2, t_3, \dots, t_m\}$

1. Set $R_{up} \leftarrow \{root(t_1), root(t_2), \dots, root(t_m)\}$
2. Set $R_p \leftarrow \emptyset$
3. while $\{R_{up} \neq \emptyset\}$
 - do
 - 3.1 Pick an r from R_{up}
 - 3.2 If (r forms a cycle with an r' in R_p) then
 - 3.2.1 Delete an edge e_r incident on r and an edge $e_{r'}$ incident on r'
 - 3.2.2 Add roots of the subtrees of r and r' to R_{up}
 - 3.2.3 Continue
 - 3.3 else $R_p \leftarrow R_p + r$
- od
4. Return the trees whose roots are in R_p

2-approximation ratio claim

- The lemma below justifies the choice of edges removed in MAAF-Approx and aids in the analysis of the approximation ratio algorithm

Lemma 6 *If $e(F, T_2, \dots, T_k)$ is the minimum number of edges that must be removed from F to obtain a MAAF, then $e(F - \{e_x\}, T_2, \dots, T_k) = e(F, T_2, \dots, T_k) - 1$, where $x \in r, r'$; moreover, $e(F - \{e_r, e_{r'}\}, T_2, \dots, T_k) \leq e(F, T_2, \dots, T_k) - 1$*

2-Approximation claim

- As in the analysis of the Approx-MAF algorithm ,we can show the following:
 - $\beta' \leq e(F_A) \leq 2\beta'$, where β' is the number of iterations of the Approx-MAAF
 - From this the 2-approximation claim follows

Carrying On....

- Find approximation ratios for computing MAF and MAAF on k unrooted trees
- Extend this work to k trees of degree d ($d \geq 2$)
- Test with real biological datasets
- Explore Fixed-Parameter Tractability approach to computing MAF and MAAF on k rooted phylogeny trees.

References

- Allen, B.L., Steel, M. 2001. Subtree Transfer Operations and their Induced Metrics on Evolutionary Trees. *Annals of Combinatorics*. 1-15.
- Baroni, M., Grunewald, S., Moulton, V., Semple, C. 2005. Bounding the number of Hybridisation Events for a Consistent Evolutionary History. *Mathematical Biology*. 51, 171-182
- Bordewich, M., McCartin, C., Semple, C. 2008. A 3-approximation algorithm for the subtree distance between phylogenies. *Journal of Discrete Algorithms*
- Chataigner, F. 2005. Approximating the Maximum Agreement Forest on k trees. *Information Processing Letters*. 93, 239-244
- Hein J., Jiang, T., Wang, L., Zhang, K. 1996. On the complexity of comparing evolutionary trees. *Discrete Appl. Math.* 71, 153–169.
- Rodrigues, E.M., Sagot, M.F., Wakabayashi, Y. 2007. The Maximum agreement forest problem: Approximation algorithms and computational experiments. *Theoretical Computer Science*, 374(1-3):91-110
- Whidden, C., Zeh, N. 2009. A Unifying View on Approximation and FPT of Agreement Forests. *WABI*. 5724: 390-402
- Wu, Y., Wang, J. 2010. Fast Computation of the Exact Hybridization Number of two phylogenetic trees. *ISBRA*. 203-214
- Inaugural-Dissertation on Reticulation in Evolution from Simone Linz

Thank You!

