Phylogenetics meets Courcelle's Theorem

Steven Kelk

Department of Data Science and Knowledge Engineering (DKE)
Maastricht University, Netherlands

Spring Phylogenetics Workshop, UEA, 10th May 2016

Based on joint work with Celine Scornavacca, Leo van Iersel and Mathias Weller.

- Phylogenetic trees summarise the evolution of a set of species *X*.

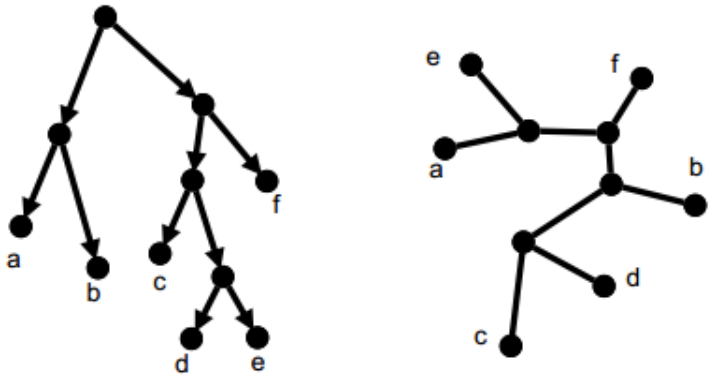
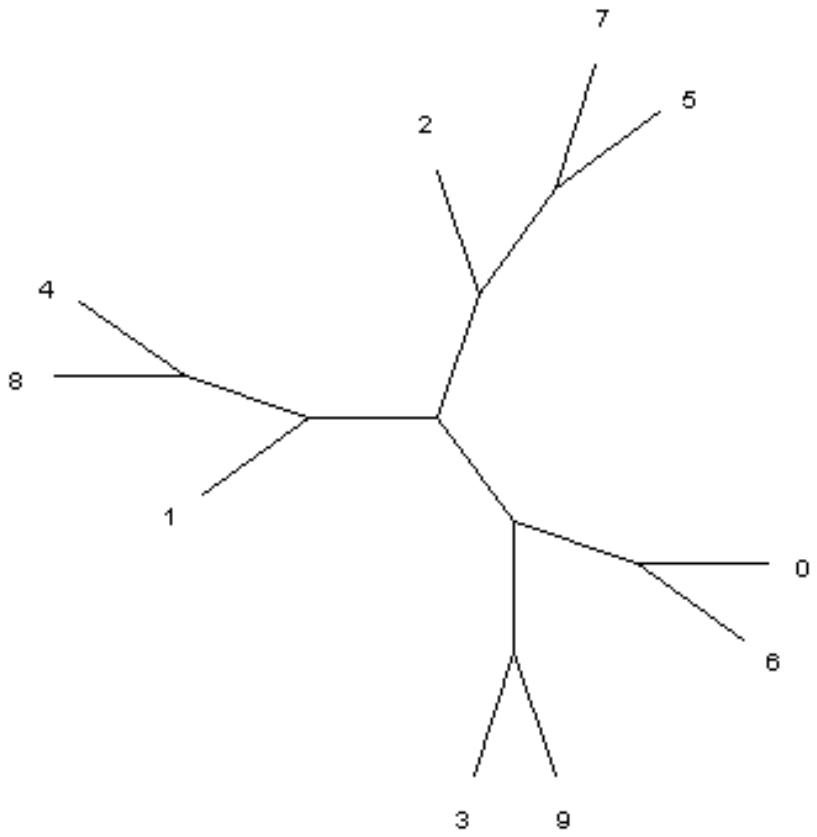


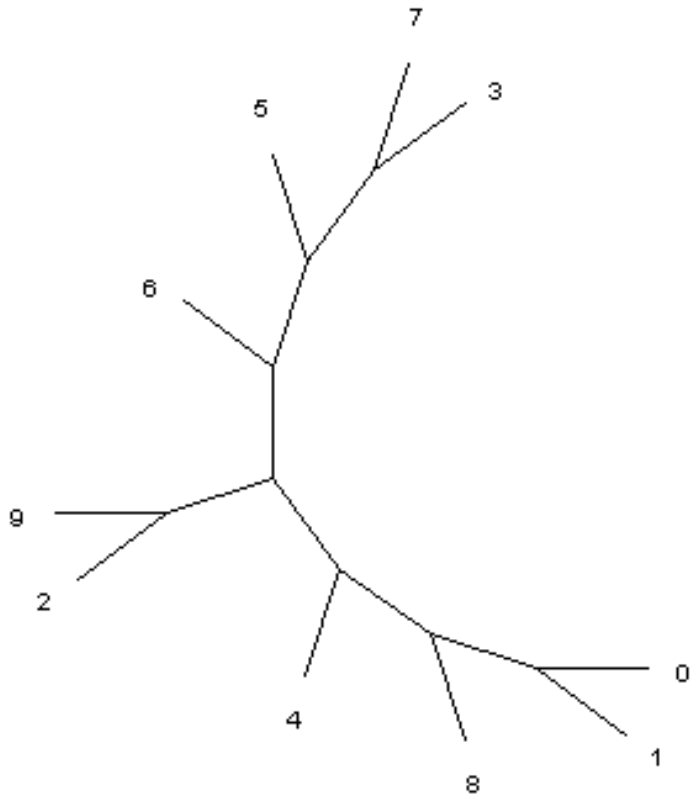
Figure 1: *(a) A rooted phylogenetic tree on* $X = \{a, \ldots, f\}$. *(b) An unrooted phylogenetic tree on the same set* X*; note that the edges are undirected and there is no root. Both trees are binary.*

• The central goal of phylogenetics is to *infer* these trees from e.g. DNA data.

• However, phylogenetics software often generates several topologically distinct (*“incongruent”*) trees.

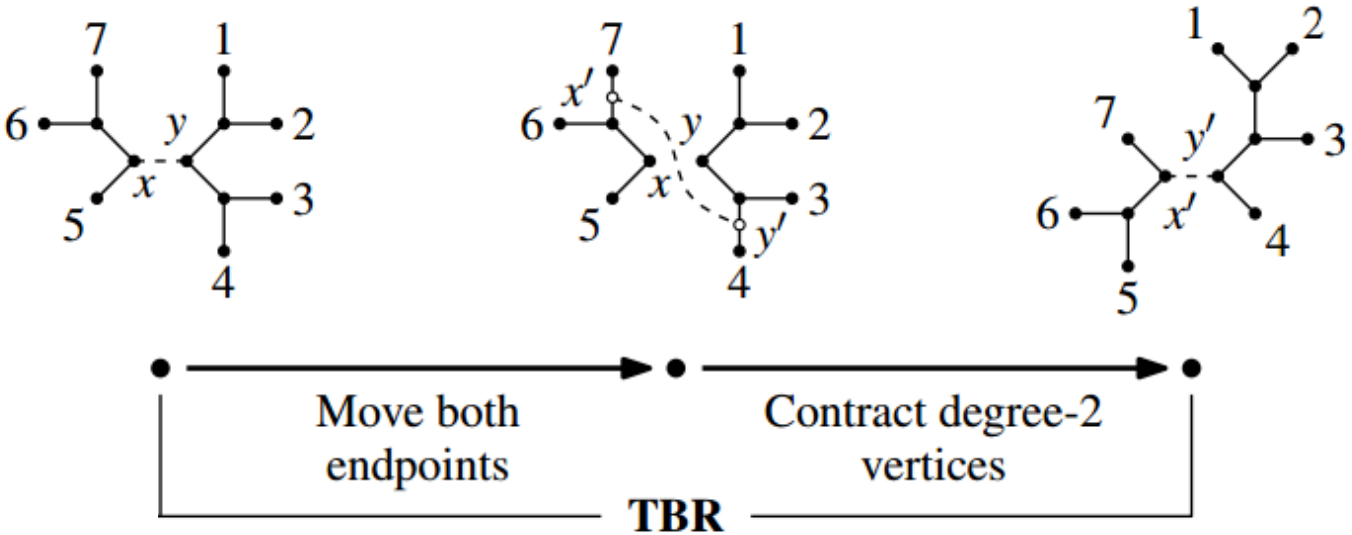
• Important to quantify incongruence i.e. in how far two (or more) trees differ from each other topologically.


7
5
2
4
8
1
0
6
3
9


7
3
5
6
9
2
0
4
1
8

- Lots of NP-hard problems!

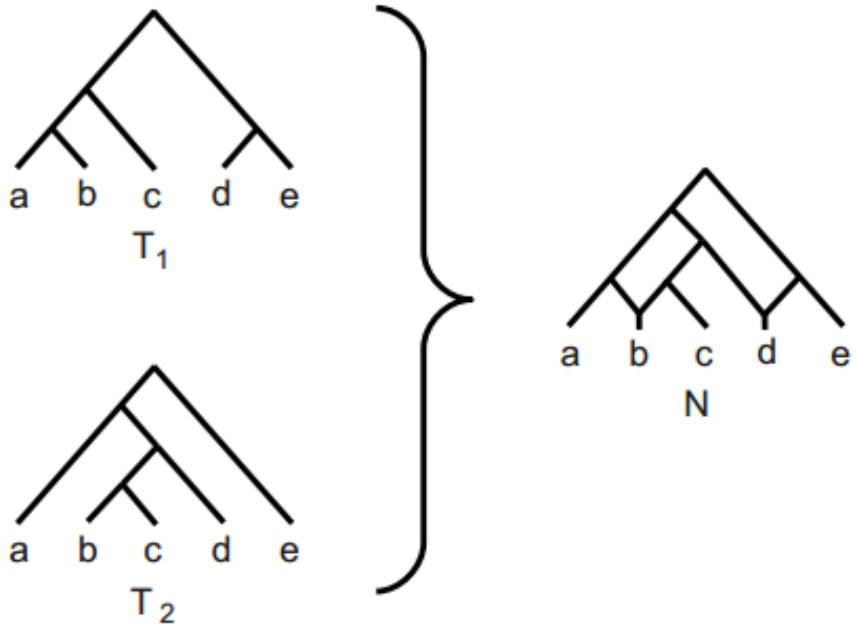
- Two main approaches:
  - Distance measures – minimizing number of rearrangement "moves" to transform one tree into another, e.g. TBR.



From Whidden & Zeh 2009

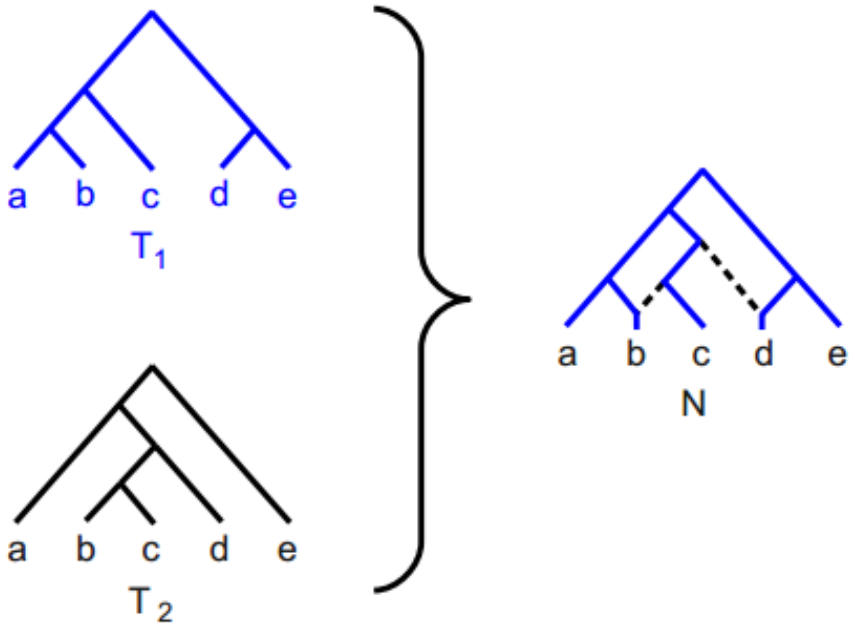
- Lots of NP-hard problems!

- Two main approaches:
  - Constructive approaches: merging incongruent (rooted) trees into a phylogenetic network: a DAG that summarizes all trees



- Lots of NP-hard problems!

- Two main approaches:
  - Constructive approaches: merging incongruent (rooted) trees into a phylogenetic network: a DAG that summarizes all trees



- Many FPT results: running times of the form *f(k) * poly(n)*, where *k* is (usually) the phylogenetic “distance” measure we are computing

- Kernelization, bounded search

- Literature isolated from rest of algorithmic graph theory

- Modelling is complex!

- Problems do not have the classical form i.e.
  - Input: A graph G
  - Output: A minimum size vertex cover

- Can we close this modelling barrier?

- Many FPT results: running times of the form *f(k) * poly(n)*, where *k* is (usually) the phylogenetic “distance” measure we are computing

- Kernelization, bounded search

- Literature isolated from rest of algorithmic graph theory

- Modelling is complex!

- Problems do not have the classical form i.e.
  - Input: A graph G
  - Output: A minimum size vertex cover

- Can we close this modelling barrier?

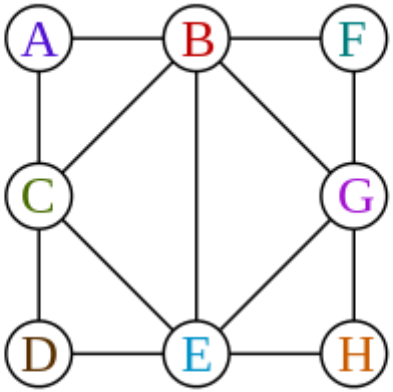
- **Why** do we want to close this modelling barrier?

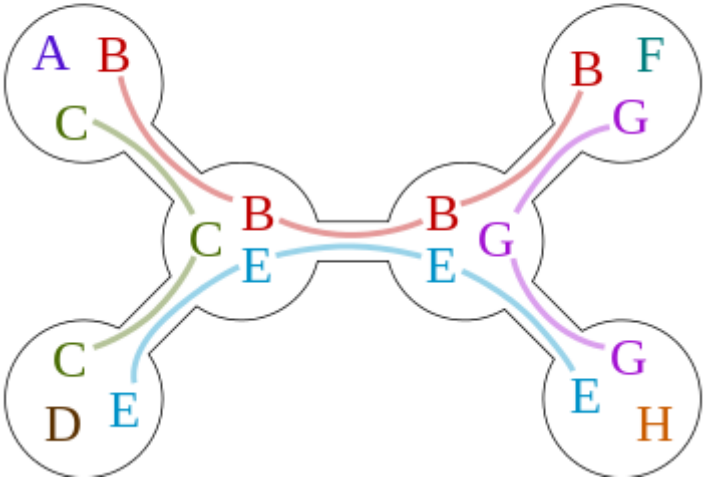
• The concept of **treewidth** lies at the foundation of many deep, structural results in algorithmic graph theory.

• Informally, treewidth measures “how far” an undirected graph is from being a tree. (Trees have treewidth 1).

• Definition is quite technical.

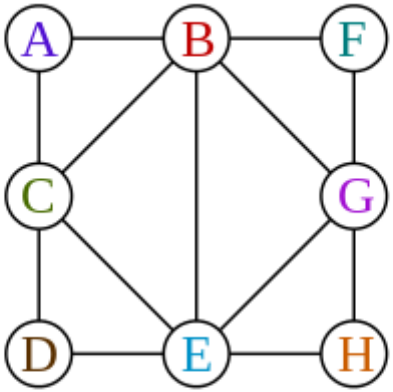
• This graph has treewidth 2.

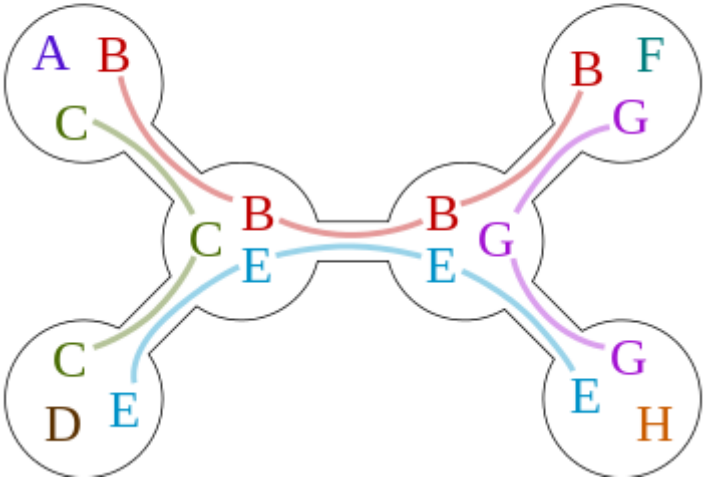




Source: *wikipedia*

• Very many NP-hard problems can be solved in time *f(t) * poly(n)* on graphs of treewidth at most *t,* using dynamic programming

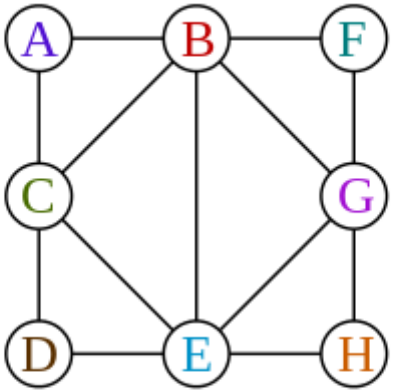


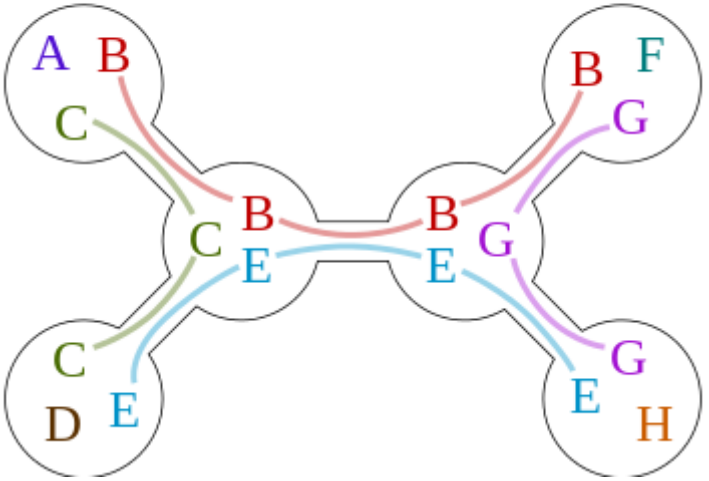


Source: *wikipedia*

• Very many NP-hard problems can be solved in time *f(t) * poly(n)* on graphs of treewidth at most *t,* using dynamic programming

•This stimulated an abstraction / generalization process, leading to the emergence of **algorithmic meta-theorems**

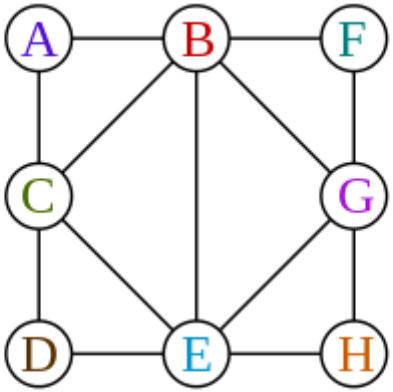


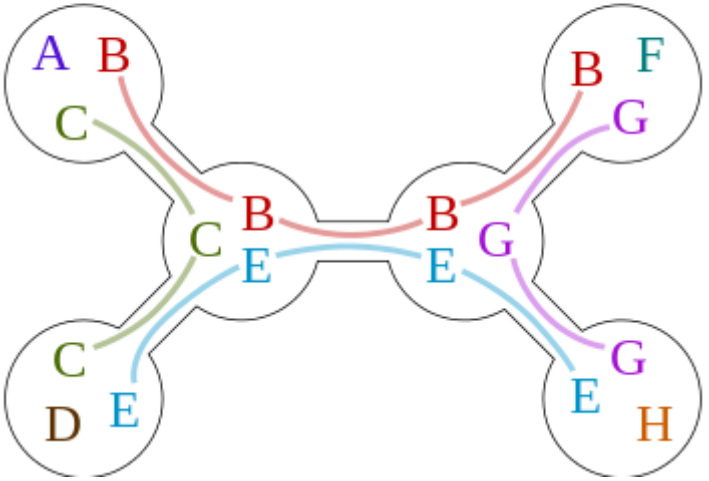


Source: *wikipedia*

• **Courcelle's Theorem** (Courcelle 1990, Arnborg et al 1991)

• Any property of an undirected graph that can be expressed as a length *L* fragment of monadic second order logic, can be answered in time *f(L,t)* * *O(n)* on graphs of treewidth at most *t*.



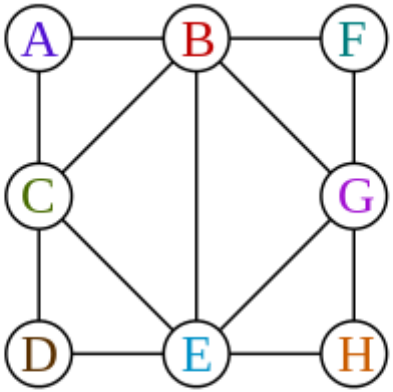


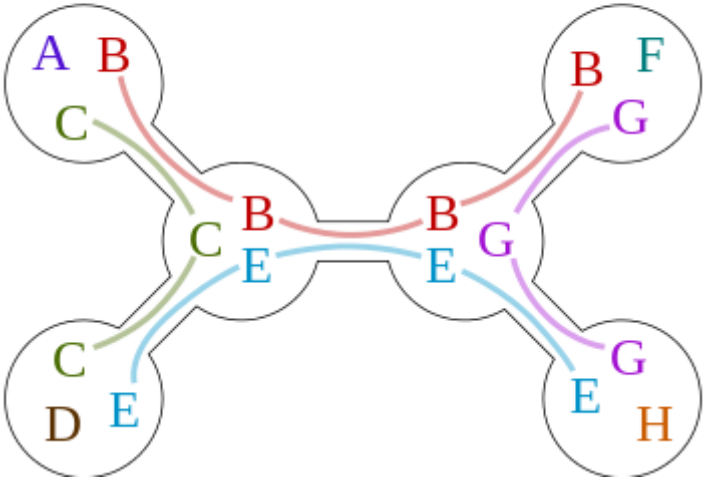
Source: *wikipedia*

• **Courcelle's Theorem** (Courcelle 1990, Arnborg et al 1991)

• Any property of an undirected graph that can be expressed as a length *L* fragment of monadic second order logic, can be answered in time *f(L,t)* * *O(n)* on graphs of treewidth at most *t*.

• Rapid prototyping of FPT results





Source: *wikipedia*

As an example, the property of a graph being colorable with three colors (represented by three sets of vertices R, G, and B) may be defined by the monadic second-order formula

$$\exists R,G,B.\ (\forall v \in V.\ (v \in R \vee v \in G \vee v \in B)) \wedge$$
$$(\forall u,v \in V.\ ((u \in R \wedge v \in R) \vee (u \in G \wedge v \in G) \vee (u \in B \wedge v \in B)) \rightarrow \neg \mathrm{adj}(u,v)).$$

The first part of this formula ensures that the three color classes cover all the vertices of the graph, and the second ensures that they each form an independent set. (It would also be possible to add clauses to the formula to ensure that the three color classes are disjoint, but this makes no difference to the result.) Thus, by Courcelle's theorem, 3-colorability of graphs of bounded treewidth may be tested in linear time.

Source: *wikipedia*

- Can we use Courcelle's Theorem in phylogenetics?

- Unfortunately, problems do not have the classical form i.e.
  - Input: A graph G
  - Output: A minimum size vertex cover

- How to deal with the fact that in phylogenetics the input is often a set of phylogenetic trees, rather than an undirected graph?

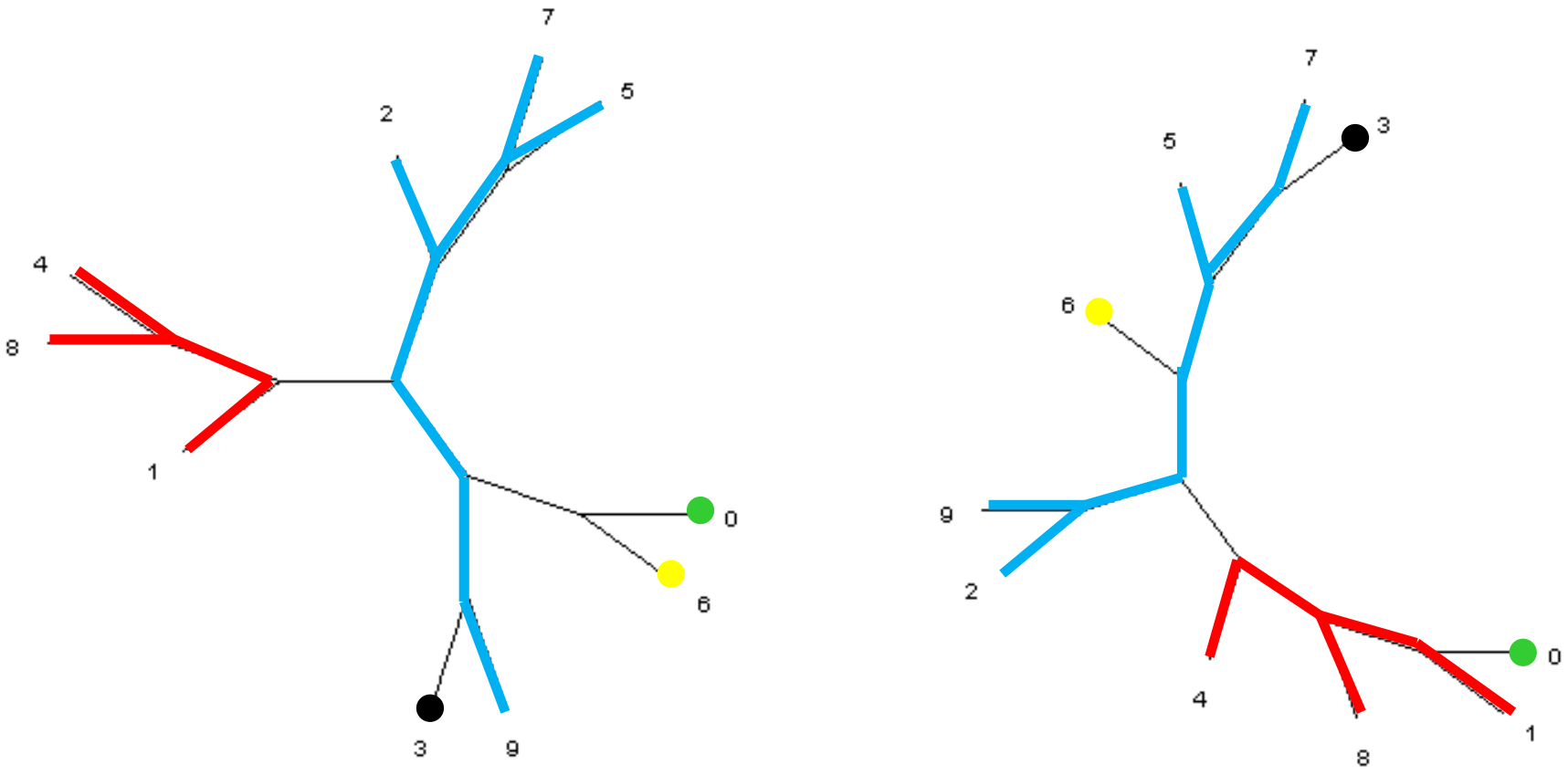
• New insights:

• Many incongruence problems use an abstraction known as an **agreement forest**

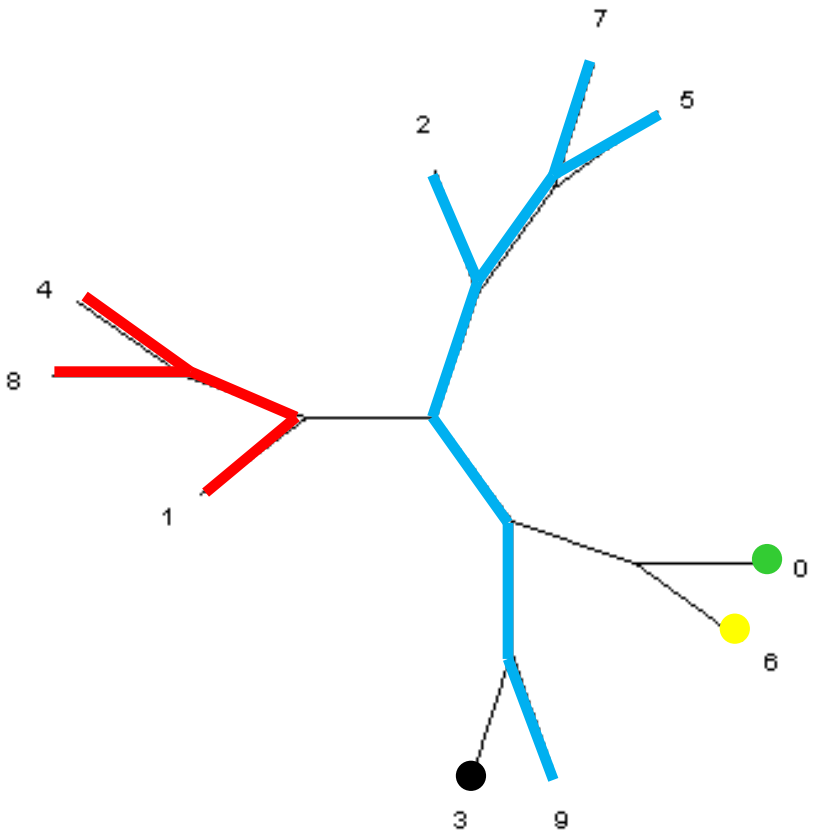
• Agreement forests induce bounded treewidth in an auxiliary graph structure known as the **display graph** (introduced by Bryant and Lagergren)

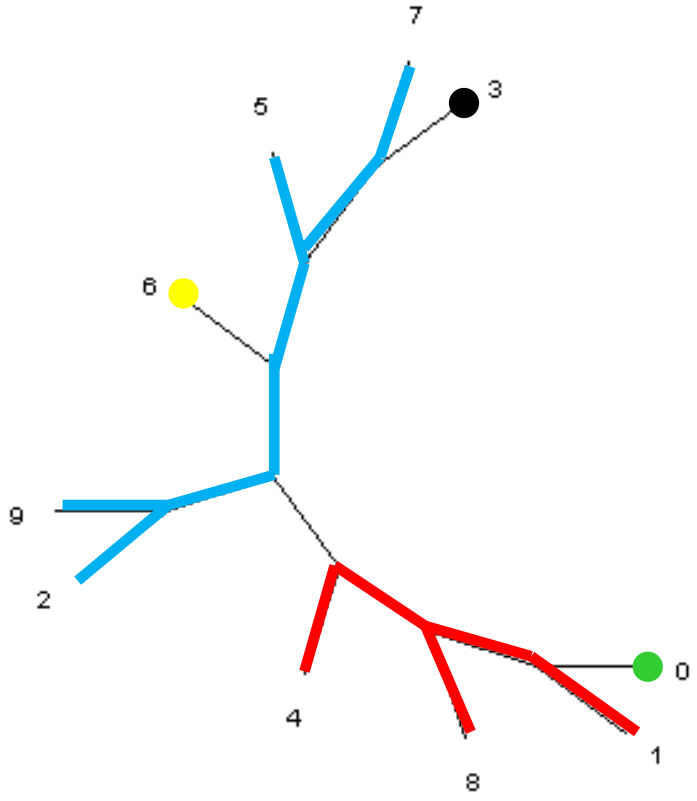
• Rapid prototyping of FPT results using (somewhat exotic…) MSO formulations on the display graph, some using the treewidth bound constructively, others only implicitly

• Proof of concept

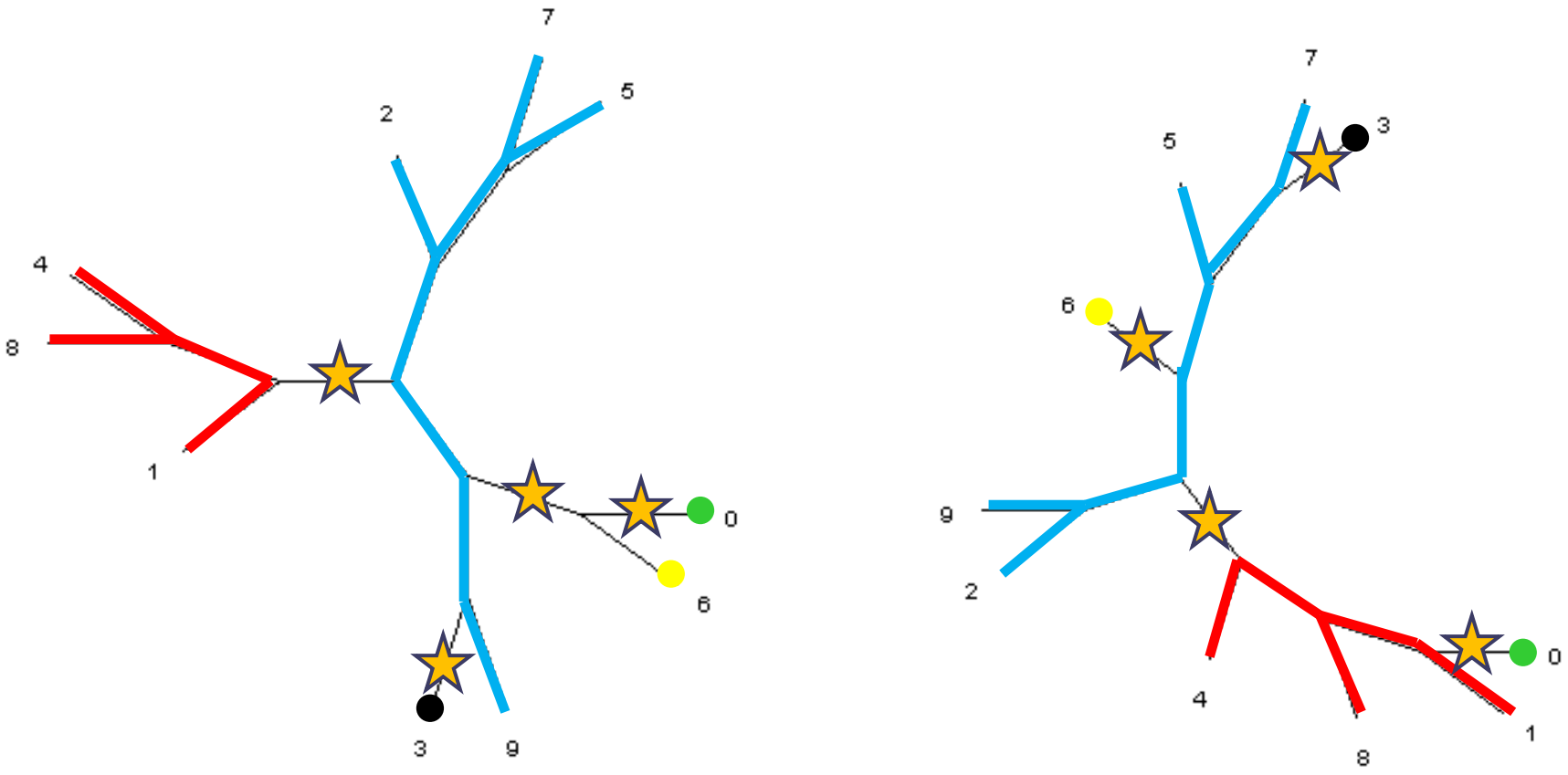


Agreement forest with 5 components

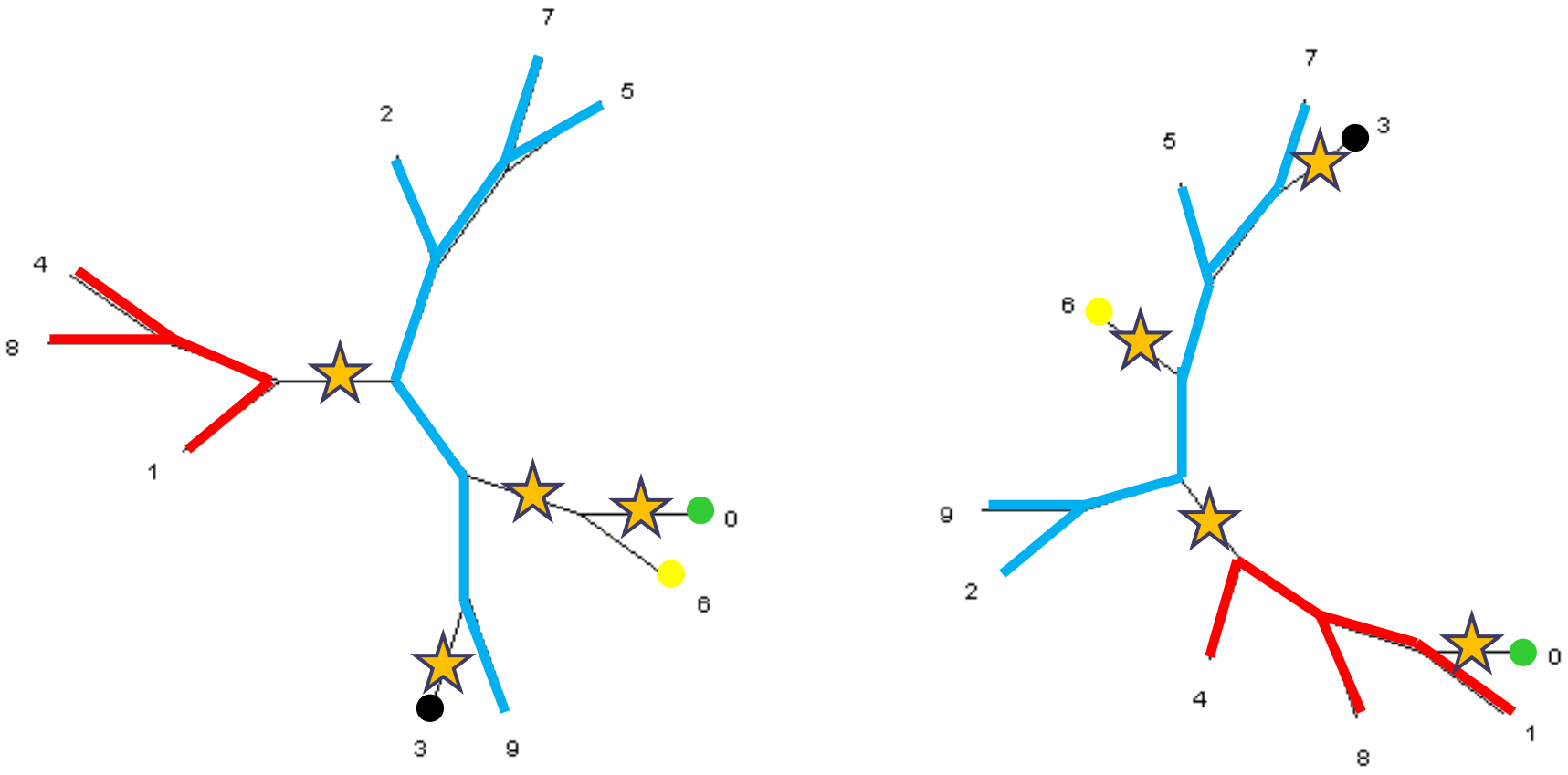




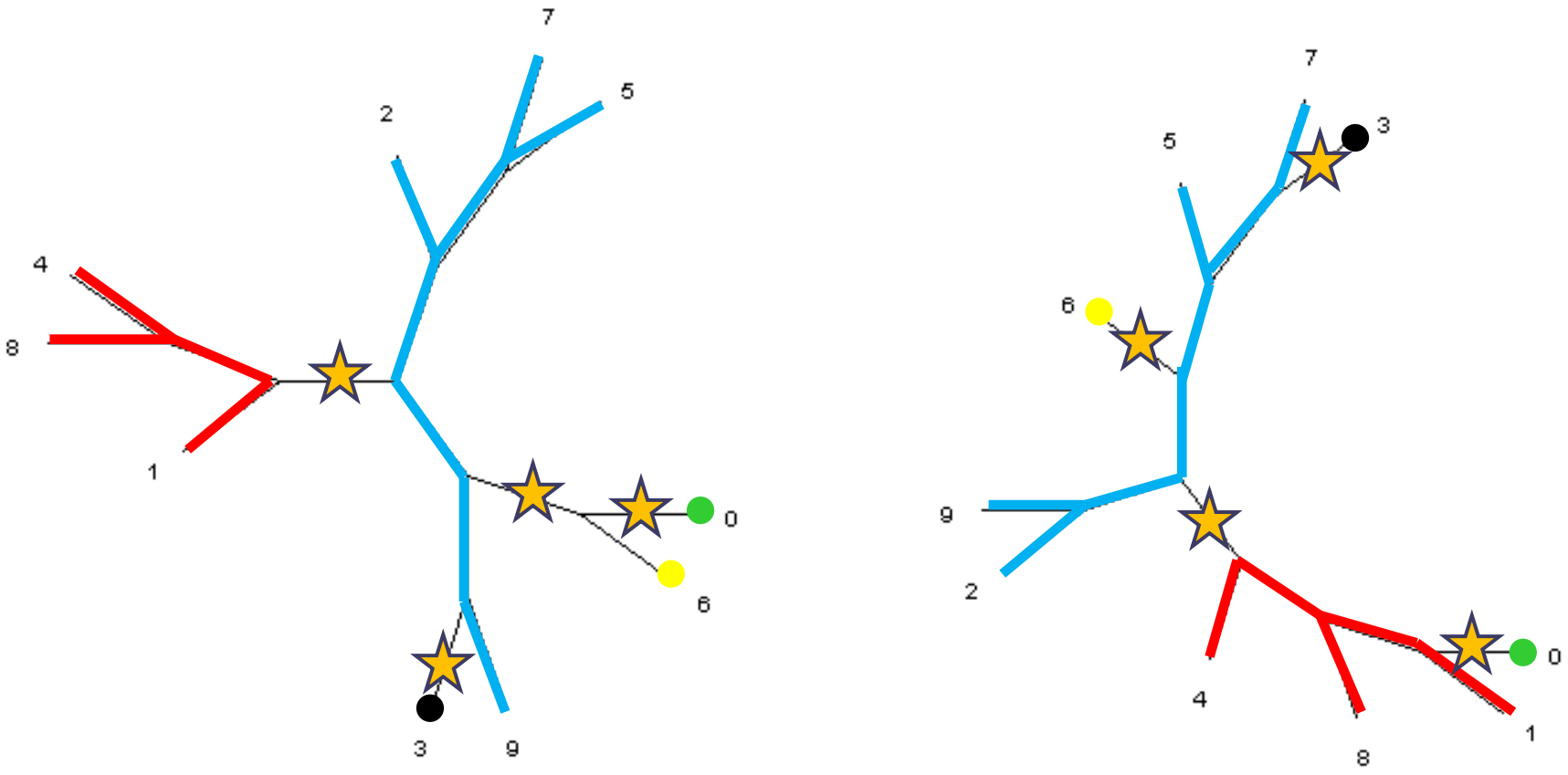
Obtained by cutting 4 edges in each tree



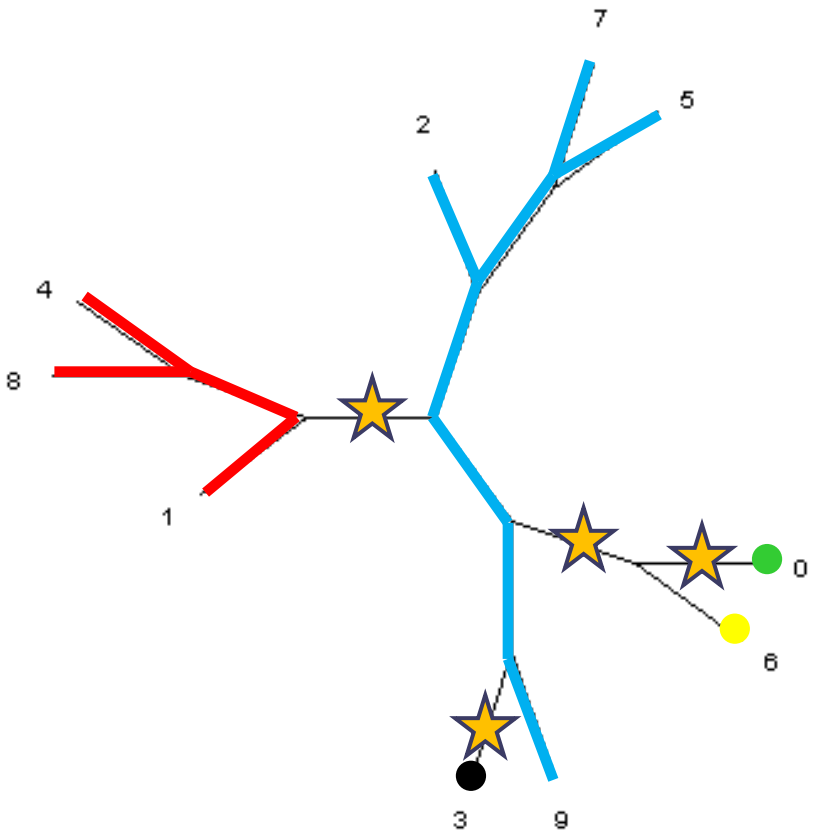
Obtained by cutting 4 edges in each tree

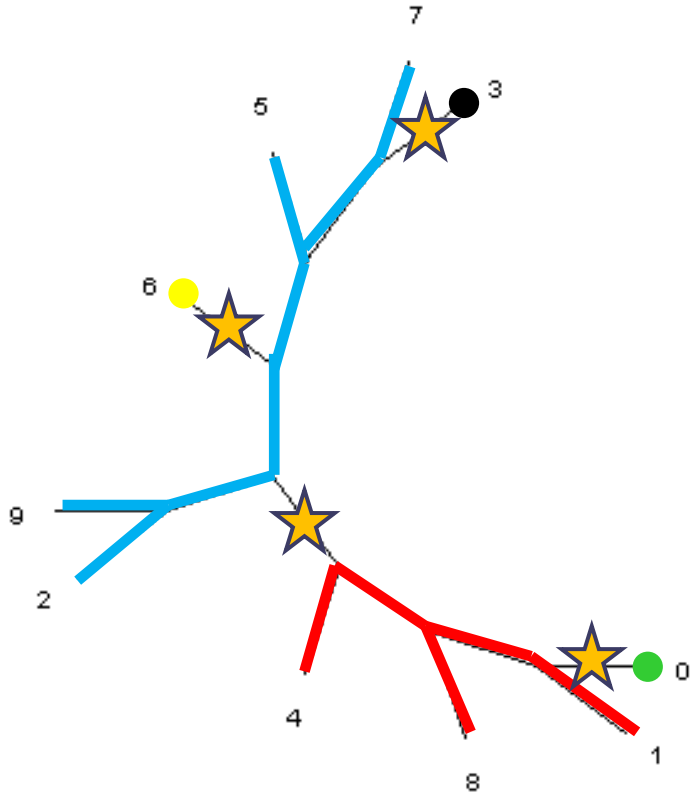


Fewer components are not possible: this is a Maximum Agreement forest (MAF)



Allen & Steel 2001:
TBR distance = #components in MAF - 1





Allen & Steel 2001:
TBR distance = 5 – 1 = 4

•The **display graph *D*** of a set of phylogenetic trees is obtained by identifying all leaves that have the same label

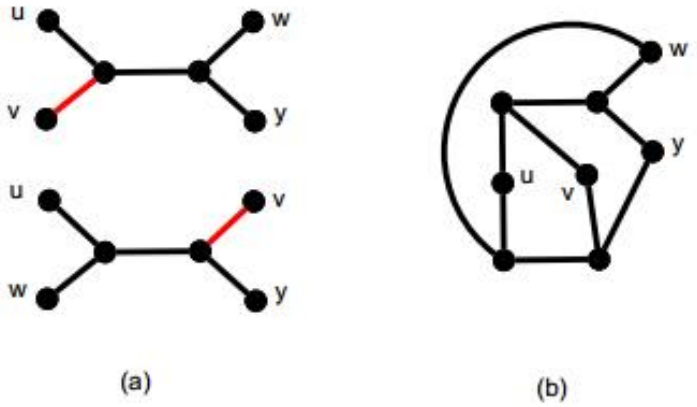


Figure 2: *(a) Two unrooted binary phylogenetic trees on* $\{u, v, w, y\}$. *A maximum agreement forest (MAF) for these two trees contains 2 components, and can be obtained by cutting the single red edge in both trees and then suppressing the resulting degree 2 vertices. (b) The display graph for the two trees from (a), obtained by identifying leaves with the same label.*

•The **display graph *D*** of a set of phylogenetic trees is obtained by identifying all leaves that have the same label

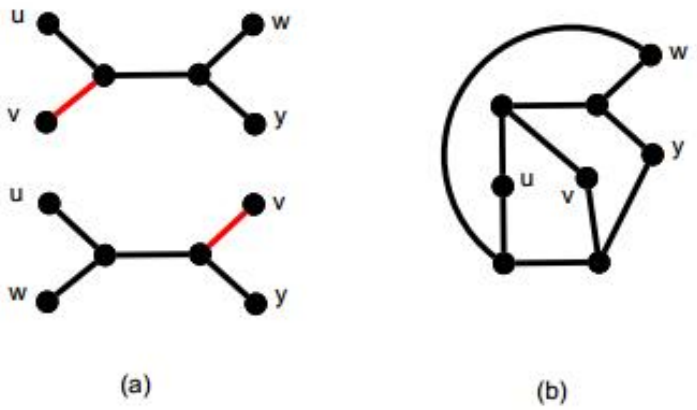
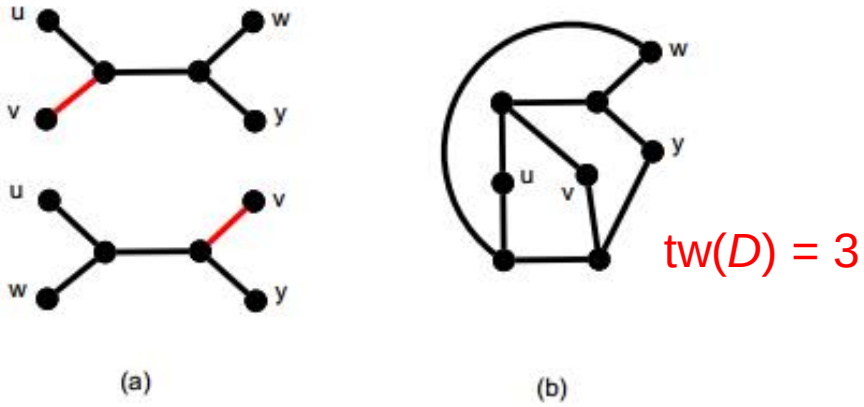


Figure 2: *(a) Two unrooted binary phylogenetic trees on* $\{u, v, w, y\}$. *A maximum agreement forest (MAF) for these two trees contains 2 components, and can be obtained by cutting the single red edge in both trees and then suppressing the resulting degree 2 vertices. (b) The display graph for the two trees from (a), obtained by identifying leaves with the same label.*

Simple but powerful insight:
tw(*D*) ≤ #MAF + 1

•The **display graph *D*** of a set of phylogenetic trees is obtained by identifying all leaves that have the same label



tw(*D*) = 3

Figure 2: *(a) Two unrooted binary phylogenetic trees on* $\{u, v, w, y\}$. *A maximum agreement forest (MAF) for these two trees contains 2 components, and can be obtained by cutting the single red edge in both trees and then suppressing the resulting degree 2 vertices. (b) The display graph for the two trees from (a), obtained by identifying leaves with the same label.*

Simple but powerful insight:
tw(*D*) ≤ #MAF + 1

• Natural MSO formulation on the display graph to query, "*Is TBR ≤ k?",* high-level idea. (Uses extended MSO framework of Arnborg et al)

• "Does there exist a set of *k-1* edge cuts in the first tree T_1, and *k-1* edge cuts in the second tree T_2, such that the induced forests are identical (i.e. is an agreement forest?)"

• Natural MSO formulation on the display graph to query, "*Is TBR ≤ k?",* high-level idea. (Uses extended MSO framework of Arnborg et al)

• "Does there exist a set of *k-1* edge cuts in the first tree T_1, and *k-1* edge cuts in the second tree T_2, such that the induced forests are identical (i.e. is an agreement forest?)"

Each forest induces the same partition of the leaf labels *X*

• Natural MSO formulation on the display graph to query, "*Is TBR ≤ k?",* high-level idea. (Uses extended MSO framework of Arnborg et al)

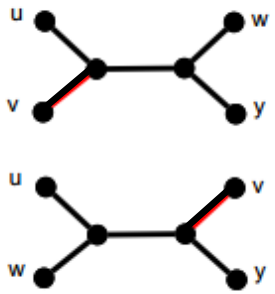
• "Does there exist a set of *k-1* edge cuts in the first tree T_1, and *k-1* edge cuts in the second tree T_2, such that the induced forests are identical (i.e. is an agreement forest?)"

Each forest induces the same partition of the leaf labels *X*

and, the forest components have the same topology

• Natural MSO formulation on the display graph to query, "*Is TBR ≤ k?*", high-level idea. (Uses extended MSO framework of Arnborg et al)

• "Does there exist a set of *k-1* edge cuts in the first tree T_1, and *k-1* edge cuts in the second tree T_2, such that the induced forests are identical (i.e. is an agreement forest?)"



Each forest induces the same partition of the leaf labels X ⇔ For each pair of leaf labels x_1 and x_2, a path from x_1 to x_2 survives the edge cuts in T_1 if and only if a path from x_1 to x_2 survives the edge cuts in T_2

and, the forest components have the same topology ⇔ The set of "quartets" that survive the edge cuts is the same in both forests

Main predicate for MAF:

$$\Big(\bigwedge_{i\in\{1,2\}} |K_i| = k' - 1\Big) \wedge \Big(\bigwedge_{i\in\{1,2\}} K_i \subseteq E_i\Big) \wedge \forall x_1, x_2 \in X(PAC(V_1, x_1, x_2, K_1) \Leftrightarrow$$
$$PAC(V_2, x_1, x_2, K_2)) \wedge \forall x_1, x_2, x_3, x_4 \in X(allDiff(x_1, x_2, x_3, x_4) \Rightarrow$$
$$((QAC^1(x_1, x_2, x_3, x_4, K_1) \Leftrightarrow QAC^2(x_1, x_2, x_3, x_4, K_2)) \wedge$$
$$(QAC^1(x_1, x_3, x_2, x_4, K_1 \Leftrightarrow QAC^2(x_1, x_3, x_2, x_4, K_2)) \wedge$$
$$(QAC^1(x_1, x_4, x_2, x_3, K_1) \Leftrightarrow QAC^2(x_1, x_4, x_2, x_3, K_2)))).$$

PAC = path avoiding cuts ?
QAC = quartet avoiding cuts ?

These auxiliary predicates operate by testing whether corresponding label-preserving topological minors survive the edge cuts

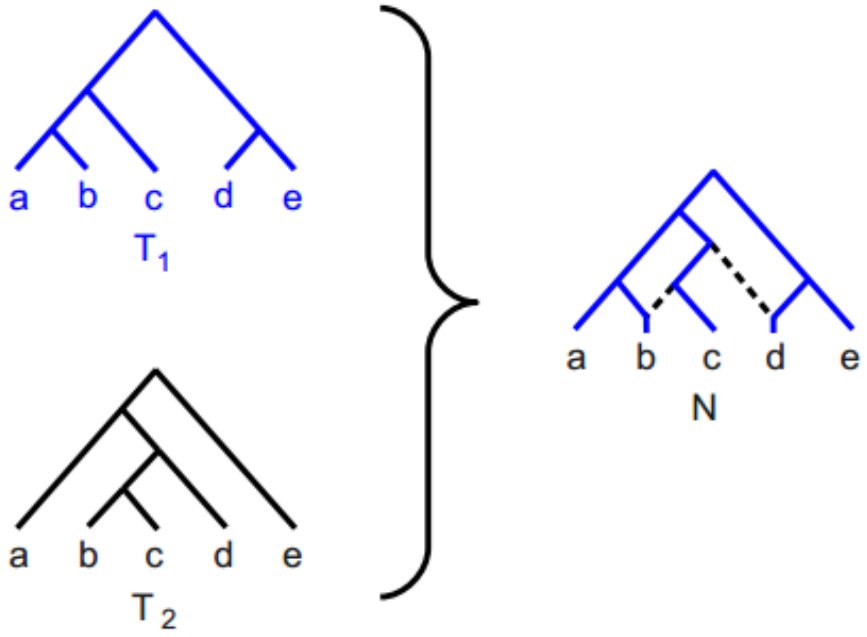
• Hence, by Courcelle's Theorem, MAF/TBR can be computed in time *f(MAF) * O(n)*

• In fact, we obtain a stronger result: it can be computed in *time f(t) * O(n)* where *t* is the treewidth of the display graph.

• **No need to explicitly describe the algorithm: logical-declarative framework for algorithm design!**

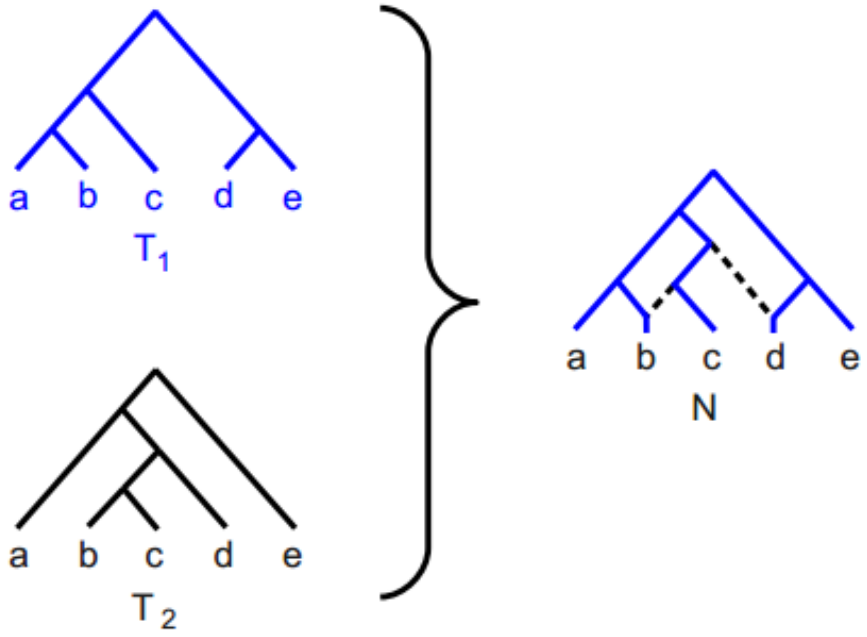
- A more exotic MSO formulation for a more exotic problem…

- *"Is Hybridization number ≤ k?"*
  - This means: Given two rooted (i.e. directed) phylogenetic trees, is there a phylogenetic network (i.e. a DAG) with at most *k* nodes of indegree 2, which contains topological embeddings of both trees?



- A more exotic MSO formulation for a more exotic problem…

- *"Is Hybridization number ≤ k?"*
  - This means: Given two rooted (i.e. directed) phylogenetic trees, is there a phylogenetic network (i.e. a DAG) with at most *k* nodes of indegree 2, which contains topological embeddings of both trees?



Here Hybridization Number = 2

• This can be modelled by imposing a special acyclicity relation on agreement forests **(Maximum Acyclic Agreement Forest, MAAF)** – Baroni et al 2005

• But tricky to directly encode this in MSO

• Idea: throw agreement forests away and recode Hybridization Number as **an elimination ordering problem**

• Essentially, the acyclicity relation is hidden within the ordering.

• MSO formulation on the display graph to query, "*Is Hybridization Number ≤ k?"* :

• *"Can we find an ordered partition of the leaf labels X into k+1 blocks, C_1, …, C_{k+1} such that, each C_i induces a common pendant subtree (CPS) assuming all C_p (p < i) have already been pruned away?"*

• *"Can we find an ordered partition of the leaf labels X into k+1 blocks, C_1, …, C_{k+1} such that, each C_i induces a common pendant subtree (CPS) assuming all C_p ($p < i$) have already been pruned away?"*

$$
\begin{aligned}
HybNum[k'](T_1, T_2) := \exists C_1, \ldots, C_{k'}, C_{k'+1} (Partition(X, C_1, \ldots, C_{k'+1}) \wedge \\
CPS(T_1, T_2, C_1, \emptyset) \wedge \\
CPS(T_1, T_2, C_2, C_1) \wedge \\
CPS(T_1, T_2, C_3, C_1, C_2) \wedge \\
\ldots \\
CPS(T_1, T_2, C_{k'+1}, C_1, \ldots, C'_k)).
\end{aligned}
$$

Main predicate

CPS is itself built from complex auxiliary predicates which enforce that the subtree in question has the same set of leaf labels and the same topology in both trees (taking into account the earlier pruning steps)

And so on…

And so on…

But is this not just abstract Theoretical Computer Science with no basis in reality?

tree pair		*taxa*	*HN*	*rSPR*	*TBR*	*uMAF*	**TW** $\leq$	*display graph size*	d^2_{MP}
rpoC2	waxy	10	1	1	1	2	**3**	$\|V\|$=28,$\|E\|$=36	1
phyB	waxy	14	3	3	2	3	**3**	$\|V\|$=40,$\|E\|$=52	2
phyB	rbcL	21	4	4	4	5	**3**	$\|V\|$=61,$\|E\|$=80	3
rbcL	waxy	12	7	6	3	4	**3**	$\|V\|$=34,$\|E\|$=44	3
phyB	rpoC2	21	7	6	4	5	**3**	$\|V\|$=61,$\|E\|$=80	3
waxy	ITS	15	8	7	5	6	**4**	$\|V\|$=43,$\|E\|$=56	3
phyB	ITS	30	8	8	7	8	**4**	$\|V\|$=88,$\|E\|$=116	5
ndhF	waxy	19	9	7	4	5	**4**	$\|V\|$=55,$\|E\|$=72	3
ndhF	rpoC2	34	12	11	8	9	**5**	$\|V\|$=100,$\|E\|$=132	6
rbcL	rpoC2	26	13	11	6	7	**5**	$\|V\|$=76,$\|E\|$=100	4
ndhF	rbcL	36	13	10	6	7	**3**	$\|V\|$=106,$\|E\|$=140	4
rbcL	ITS	29	14	13	10	11	**5**	$\|V\|$=85,$\|E\|$=112	6
ndhF	phyB	40	14	12	6	7	**3**	$\|V\|$=118,$\|E\|$=156	6
rpoC2	ITS	31	15	14	10	11	**6**	$\|V\|$=91,$\|E\|$=120	7
ndhF	ITS	46	19	19	15	16	**6**	$\|V\|$=136,$\|E\|$=180	10

Table 1: The results of our experiments with the *Poaceae* grass dataset.

tree pair		*taxa*	*HN*	*rSPR*	*TBR*	*uMAF*	**TW** $\leq$	*display graph size*	d^2_{MP}
rpoC2	waxy	10	1	1	1	2	**3**	$\|V\|$=28,$\|E\|$=36	1
phyB	waxy	14	3	3	2	3	**3**	$\|V\|$=40,$\|E\|$=52	2
phyB	rbcL	21	4	4	4	5	**3**	$\|V\|$=61,$\|E\|$=80	3
rbcL	waxy	12	7	6	3	4	**3**	$\|V\|$=34,$\|E\|$=44	3
phyB	rpoC2	21	7	6	4	5	**3**	$\|V\|$=61,$\|E\|$=80	3
waxy	ITS	15	8	7	5	6	**4**	$\|V\|$=43,$\|E\|$=56	3
phyB	ITS	30	8	8	7	8	**4**	$\|V\|$=88,$\|E\|$=116	5
ndhF	waxy	19	9	7	4	5	**4**	$\|V\|$=55,$\|E\|$=72	3
ndhF	rpoC2	34	12	11	8	9	**5**	$\|V\|$=100,$\|E\|$=132	6
rbcL	rpoC2	26	13	11	6	7	**5**	$\|V\|$=76,$\|E\|$=100	4
ndhF	rbcL	36	13	10	6	7	**3**	$\|V\|$=106,$\|E\|$=140	4
rbcL	ITS	29	14	13	10	11	**5**	$\|V\|$=85,$\|E\|$=112	6
ndhF	phyB	40	14	12	6	7	**3**	$\|V\|$=118,$\|E\|$=156	6
rpoC2	ITS	31	15	14	10	11	**6**	$\|V\|$=91,$\|E\|$=120	7
ndhF	ITS	46	19	19	15	16	**6**	$\|V\|$=136,$\|E\|$=180	10

Table 1: The results of our experiments with the *Poaceae* grass dataset.

- **Conclusion**

- **Agreement forests** induce **display graphs** with bounded treewidth

- MSO then gives FPT in parameter MAF. If we can formulate an incongruency parameter *P* in MSO, and *P* is *"larger than, or at least never significantly smaller than MAF"*, then we obtain FPT results for *P* too.

- MSO based on "phylogenetic predicates" and re-formulations:- compact, declarative proofs

- Treewidth seems to be low in practice…lots of follow-up questions!

- **Conclusion**

- **Agreement forests** induce **display graphs** with bounded treewidth

- MSO then gives FPT in parameter MAF. If we can formulate an incongruency parameter *P* in MSO, and *P* is *"larger than, or at least never significantly smaller than MAF"*, then we obtain FPT results for *P* too.

- MSO based on "phylogenetic predicates" and re-formulations:- compact, declarative proofs

- Treewidth seems to be low in practice…lots of follow-up questions!

- **Towards a structural, phylo-algorithmic graph theory!**

- **Conclusion**

- **Agreement forests** induce **display graphs** with bounded treewidth

- MSO then gives FPT in parameter MAF. If we can formulate an incongruency parameter *P* in MSO, and *P* is *"larger than, or at least never significantly smaller than MAF"*, then we obtain FPT results for *P* too.

- MSO based on "phylogenetic predicates" and re-formulations:- compact, declarative proofs

- Treewidth seems to be low in practice…lots of follow-up questions!

- **Towards a structural, phylo-algorithmic graph theory!**

- **Thank you for listening!**