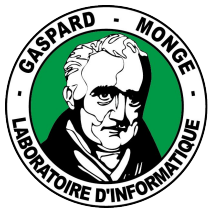
Colloquium on Networks and Evolution - 2020-09-15
Sorbonne Université (Paris)

Phylogenetic networks: how advanced are the methods?

Philippe Gambette



Outline

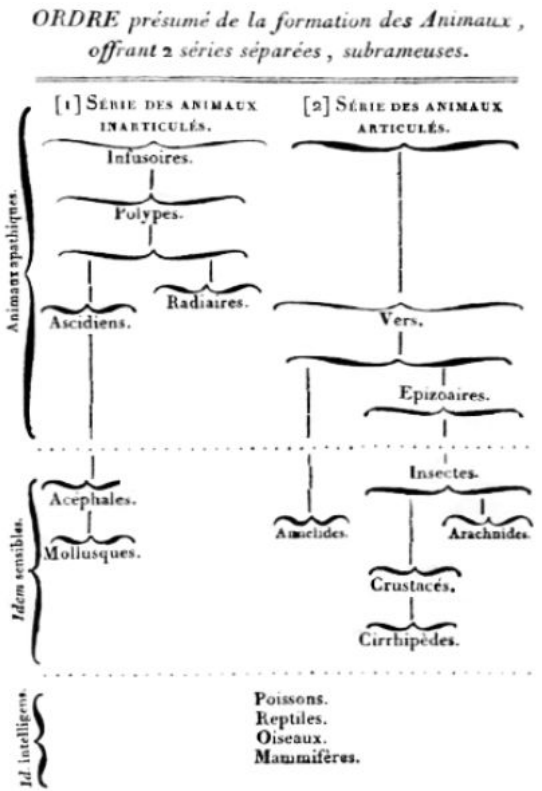
- A quick introduction to phylogenetic networks
- Advances in phylogenetic networks:
  - simplifying models
  - knowing the network space
  - finding new techniques
  - using powerful tools
  - putting everything together!

Phylogenetic networks: generalizing phylogenetic trees

Phylogenetic networks: generalizing phylogenetic trees

Phylogenetic tree of a set of species:

- **Classify them depending on common characters**
  → classification
- Describe their evolution



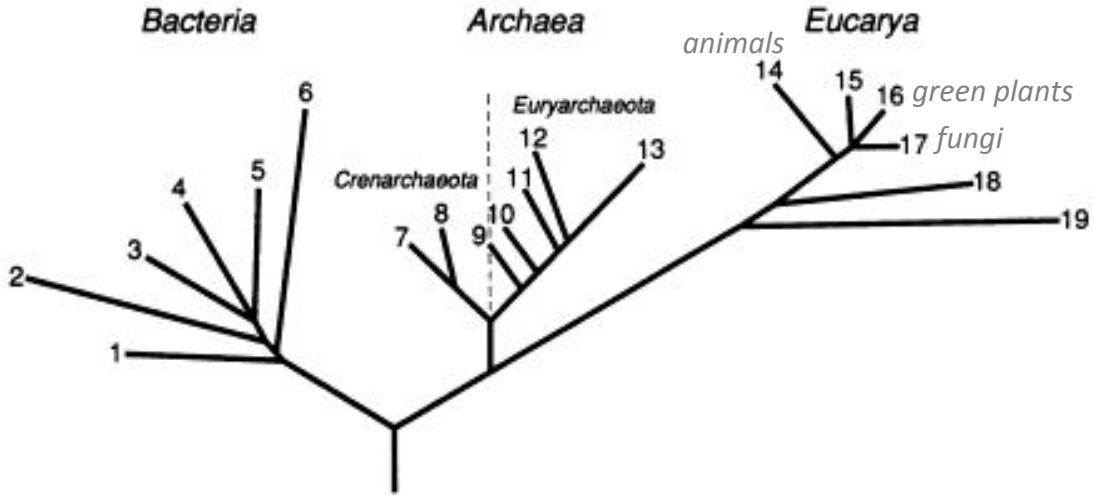
Lamarck: *Histoire naturelle des animaux sans vertèbres* (1815)

Phylogenetic networks: generalizing phylogenetic trees

Phylogenetic tree of a set of species:

- Classify them depending on common characters
- **Describe their evolution**

→ **modelization**



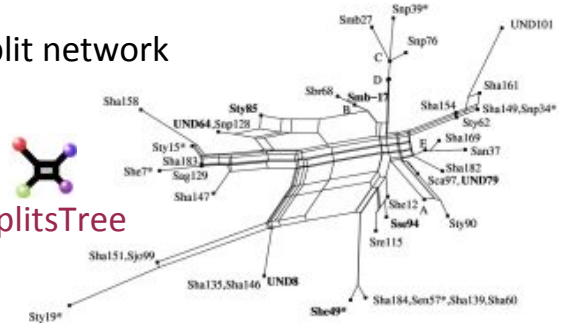
Woese, Kandler, Wheelis: Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proceedings of the National Academy of Sciences* 87(12), 4576–4579 (1990)

Abstract and explicit phylogenetic networks

Phylogenetic network: network representing evolution data

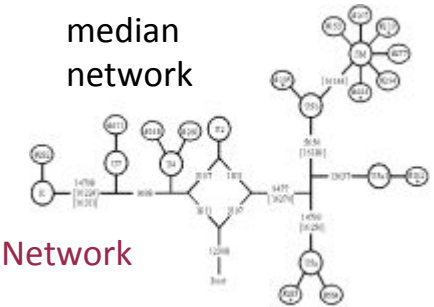
- **abstract / data-display** phylogenetic networks: to **classify**, **visualize** **data**

split network



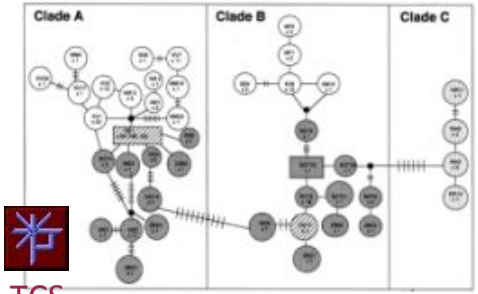
SplitsTree

median network



Network

minimum spanning network

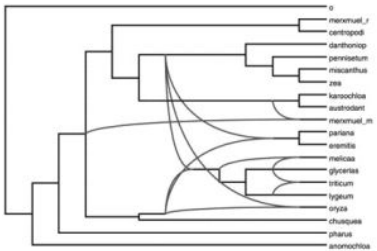


TCS

Abstract and **explicit** phylogenetic networks

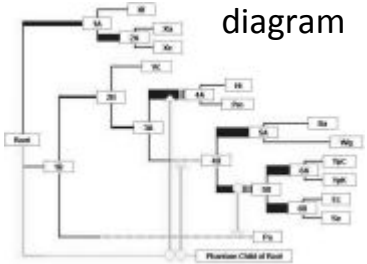
Phylogenetic network: network representing evolution data

- **explicit** phylogenetic networks: to **model** **evolution**



galled network

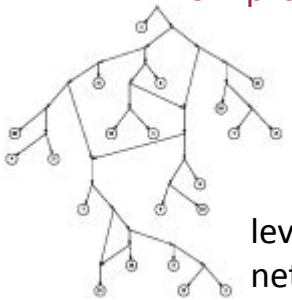
Dendroscope



synthesis diagram

HorizStory

Simplistic

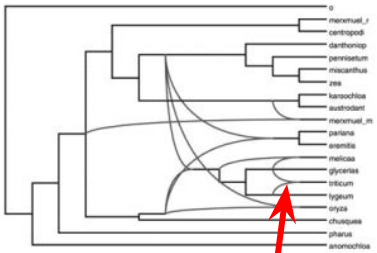


level-2 network

Abstract and **explicit** phylogenetic networks

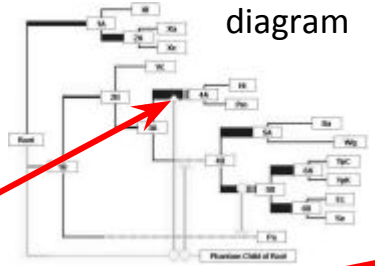
Phylogenetic network: network representing evolution data

- **explicit** phylogenetic networks: to **model** **evolution**



galled network

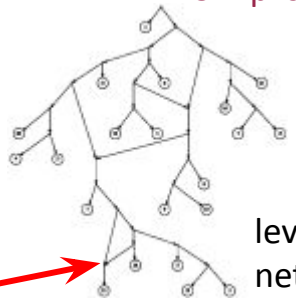
Dendroscope



synthesis diagram

HorizStory

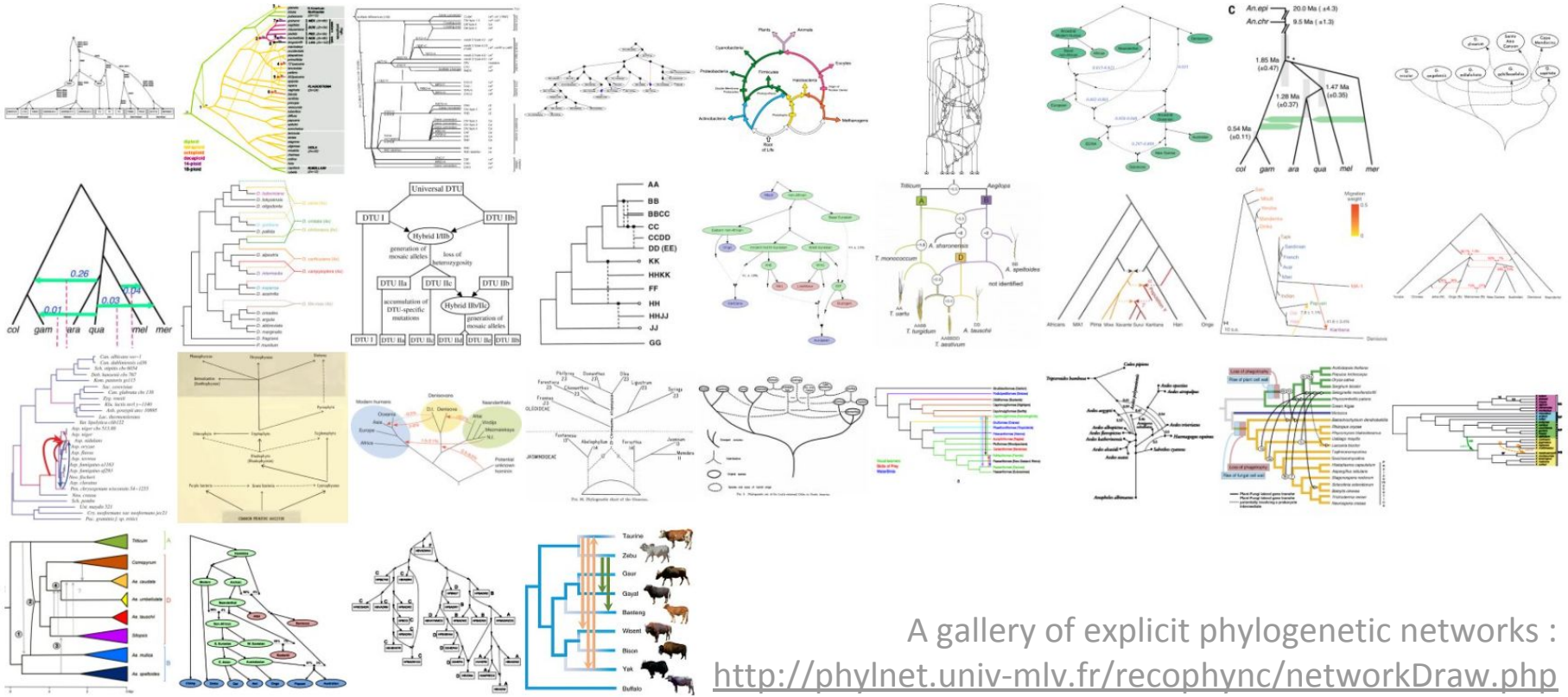
Simplistic



level-2 network

reticulations: nodes with >1 parent, modeling hybridization, recombination, lateral gene transfer, etc.

Explicit phylogenetic networks



A gallery of explicit phylogenetic networks :
http://phylnet.univ-mlv.fr/recophync/networkDraw.php

How hard is it to reconstruct a network?

How hard is it to reconstruct a network?

Quite hard.

How hard is it to reconstruct a network?

Quite hard. Harder than reconstructing trees.

How hard is it to reconstruct a network?

Quite hard. Harder than reconstructing trees. Often **NP-hard**.

How hard is it to reconstruct a network?

Quite hard. Harder than reconstructing trees. Often **NP-hard**.

In practice:

« 9.1 Limitations

The biggest limitation of methods to infer introgression and hybridization, including species network methods, is **scalability**.

Methods which infer a species network directly from multilocus sequences have only been used with **a handful of taxa**, and **less than 200 loci**. »

R. A. Leo Elworth, Huw A. Ogilvie, Jiafan Zhu and Luay Nakhleh. Advances in Computational Methods for Phylogenetic Networks in the Presence of Hybridization. In Tandy Warnow (editor), *Bioinformatics and Phylogenetics. Seminal Contributions of Bernard Moret*, Vol. 29 of *Computational Biology*, 2019

So, what should we do?

So, what should we do?
What has been done?

KEEP CALM AND SIMPLIFY YOUR MODEL

Just put gene trees together

The **“hybridization network” problem** (ignore duplication, loss, ILS, etc.):

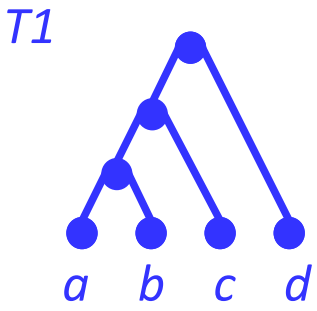
given 2 trees, find the smallest network **containing** both of them
with the minimum number of reticulations

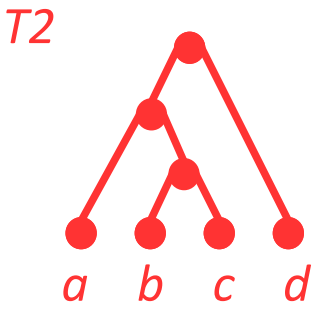
Just put gene trees together

The **“hybridization network” problem** (ignore duplication, loss, ILS, etc.):

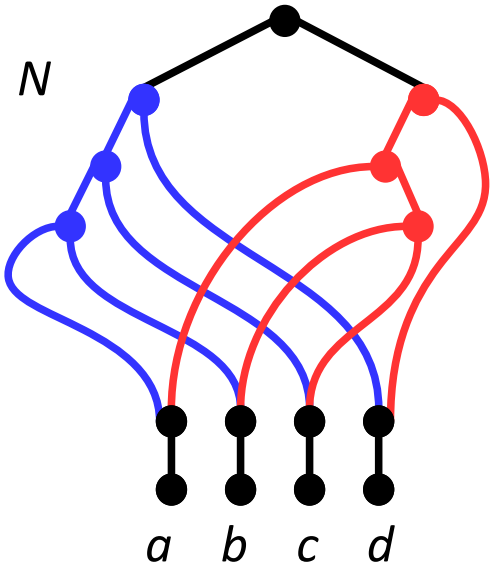
given 2 trees, find the smallest network containing both of them
with the minimum number of reticulations

Easy to find a network containing the two trees!





add a root above
the two trees, glue
the leaves together



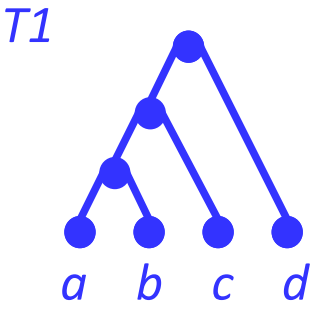
Just put gene trees together

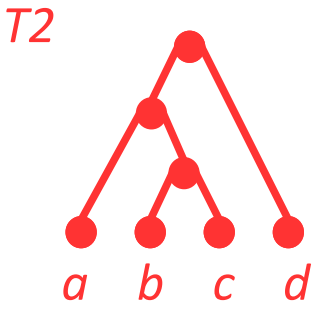
The **"hybridization network" problem** (ignore duplication, loss, ILS, etc.):

given 2 trees, find the smallest network containing both of them
with the minimum number of reticulations

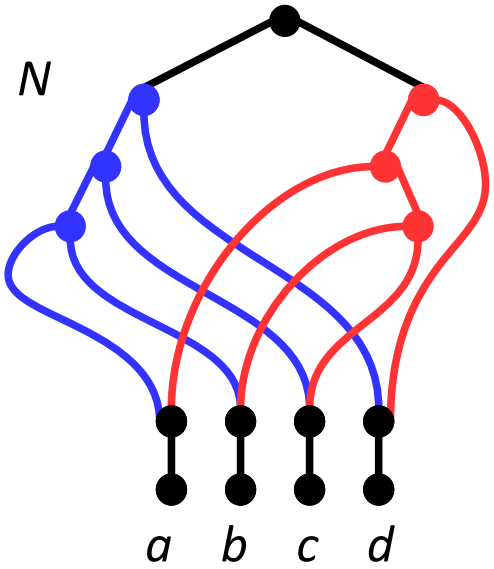
Easy to find a network containing the two trees!

But *n* hybrid vertices for trees with *n* leaves: **not optimal!**





add a root above
the two trees, glue
the leaves together



Just put gene trees together

The **“hybridization network” problem** (ignore duplication, loss, ILS, etc.):

given 2 trees, find the smallest network containing both of them
with the minimum number of reticulations

NP-hard to minimize the number of reticulations

Bordewich & Semple (2007) Discrete Appl Math

Just put gene trees together

The **"hybridization network" problem** (ignore duplication, loss, ILS, etc.):

given 2 trees, find the smallest network containing both of them
with the minimum number of reticulations

NP-hard to minimize the number of reticulations

Bordewich & Semple (2007) Discrete Appl Math

Even **checking a solution** (Tree Containment Problem) is **hard**!

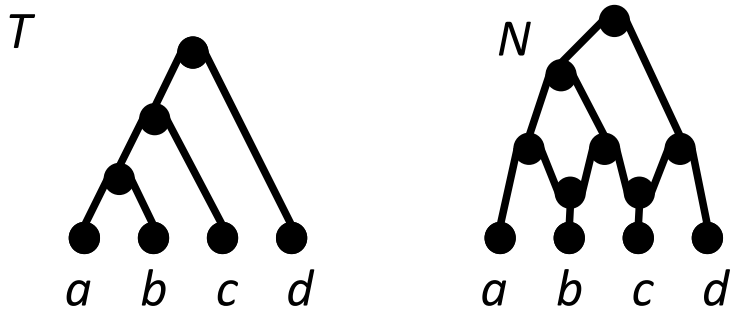
Kanj, Nakhleh, Than & Xia (2008) Theoretical Computer Science

The Tree Containment Problem

Input: A binary phylogenetic network N and a tree T over the same set of taxa.

Question: Does N display T?

→ Can we remove one incoming arc, for each vertex with >1 parent in N, so that the obtained tree is equivalent to T (each arc in T is a path in N)?

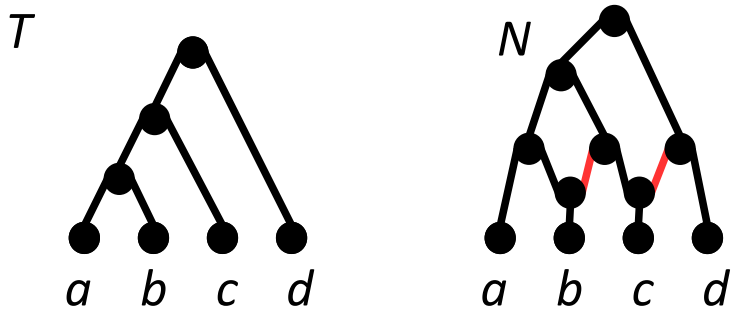


The Tree Containment Problem

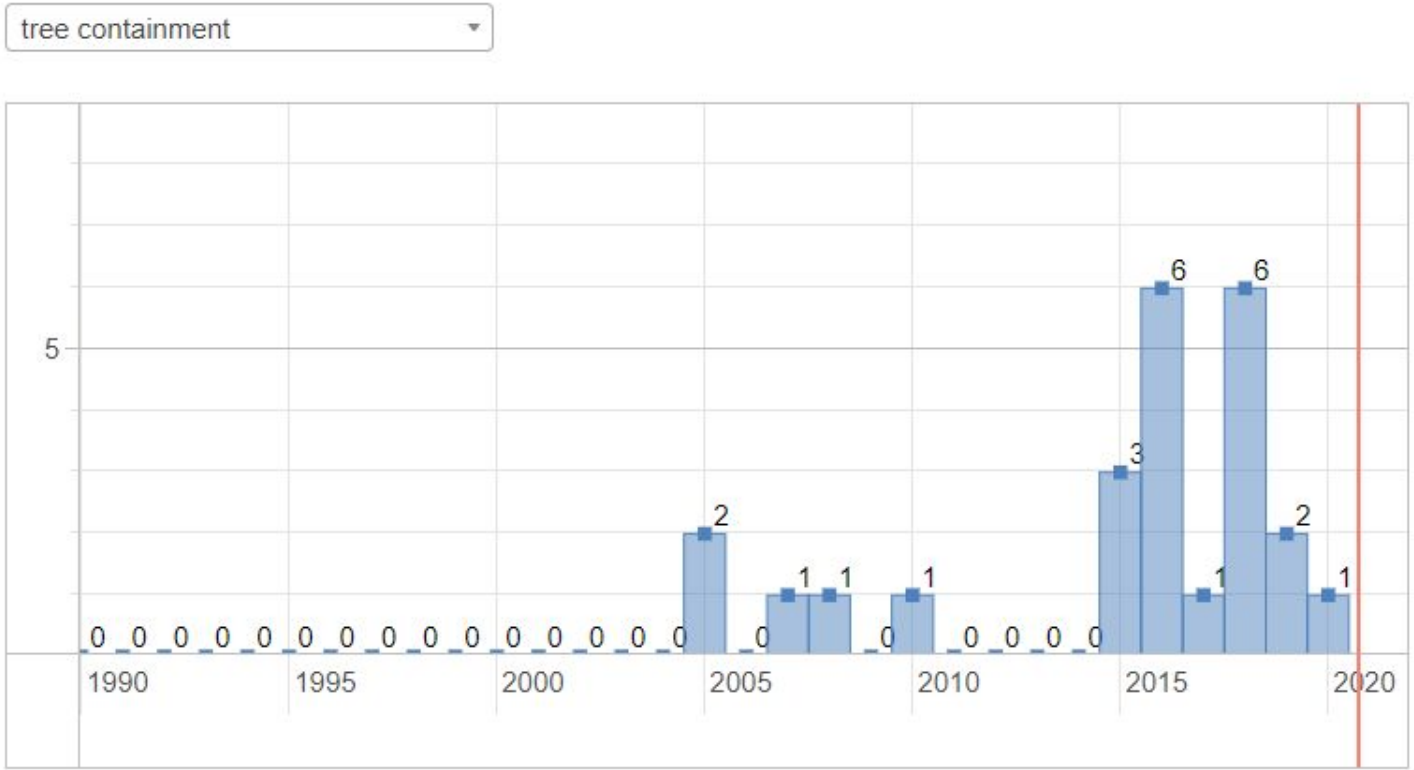
Input: A binary phylogenetic network *N* and a tree *T* over the same set of taxa.

Question: Does *N* display *T*?

→ Can we remove **one incoming arc**, for each vertex with >1 parent in *N*, so that the obtained tree is equivalent to *T* (each arc in *T* is a path in *N*)?



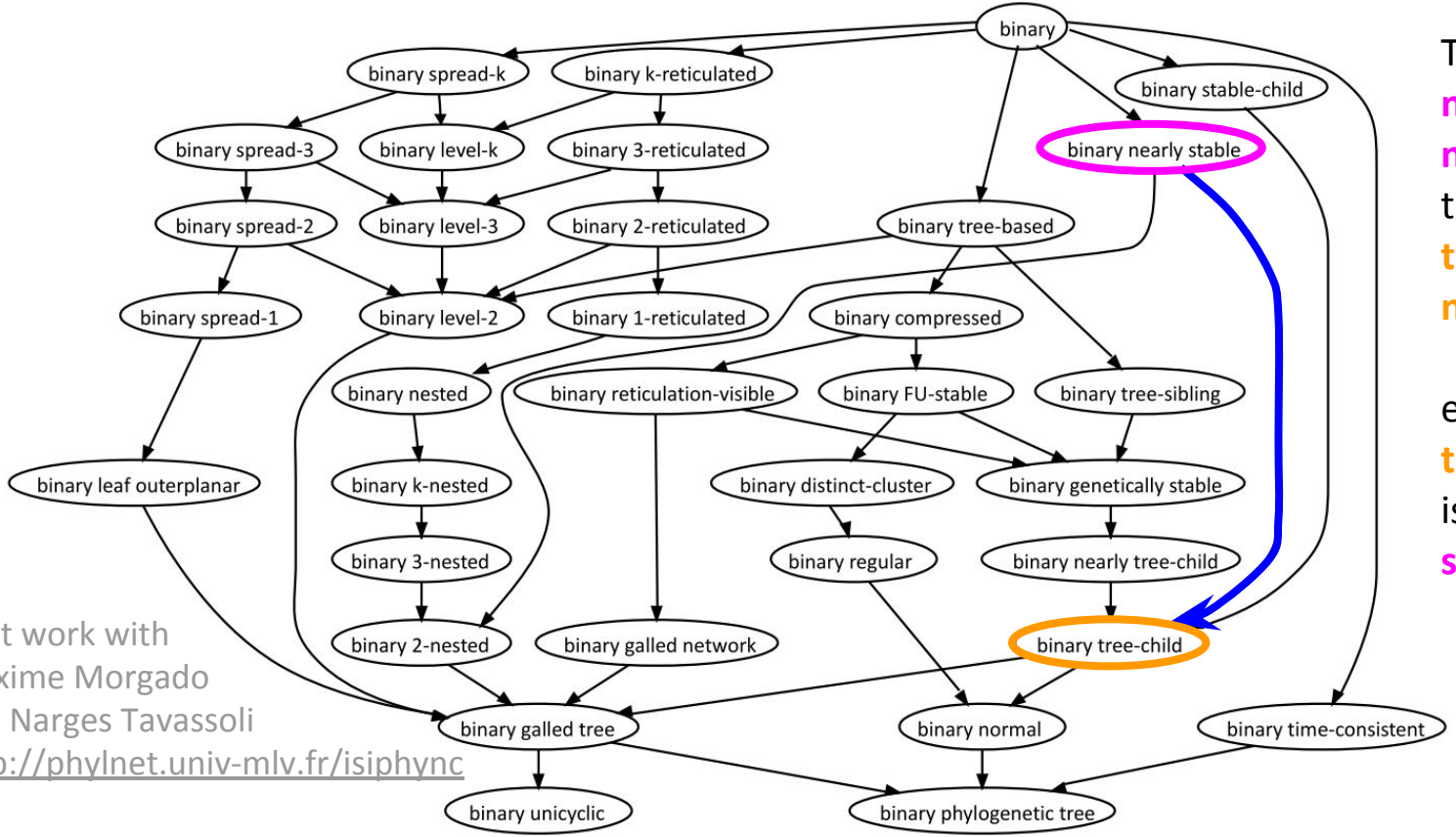
The Tree Containment Problem



Tushar Agarwal, Philippe Gambette & David Morrison (2016), *Who is Who in Phylogenetic Networks: Articles, Authors and Programs*, arXiv

http://phylnet.univ-mlv.fr/publications.php

Subclasses of phylogenetic networks

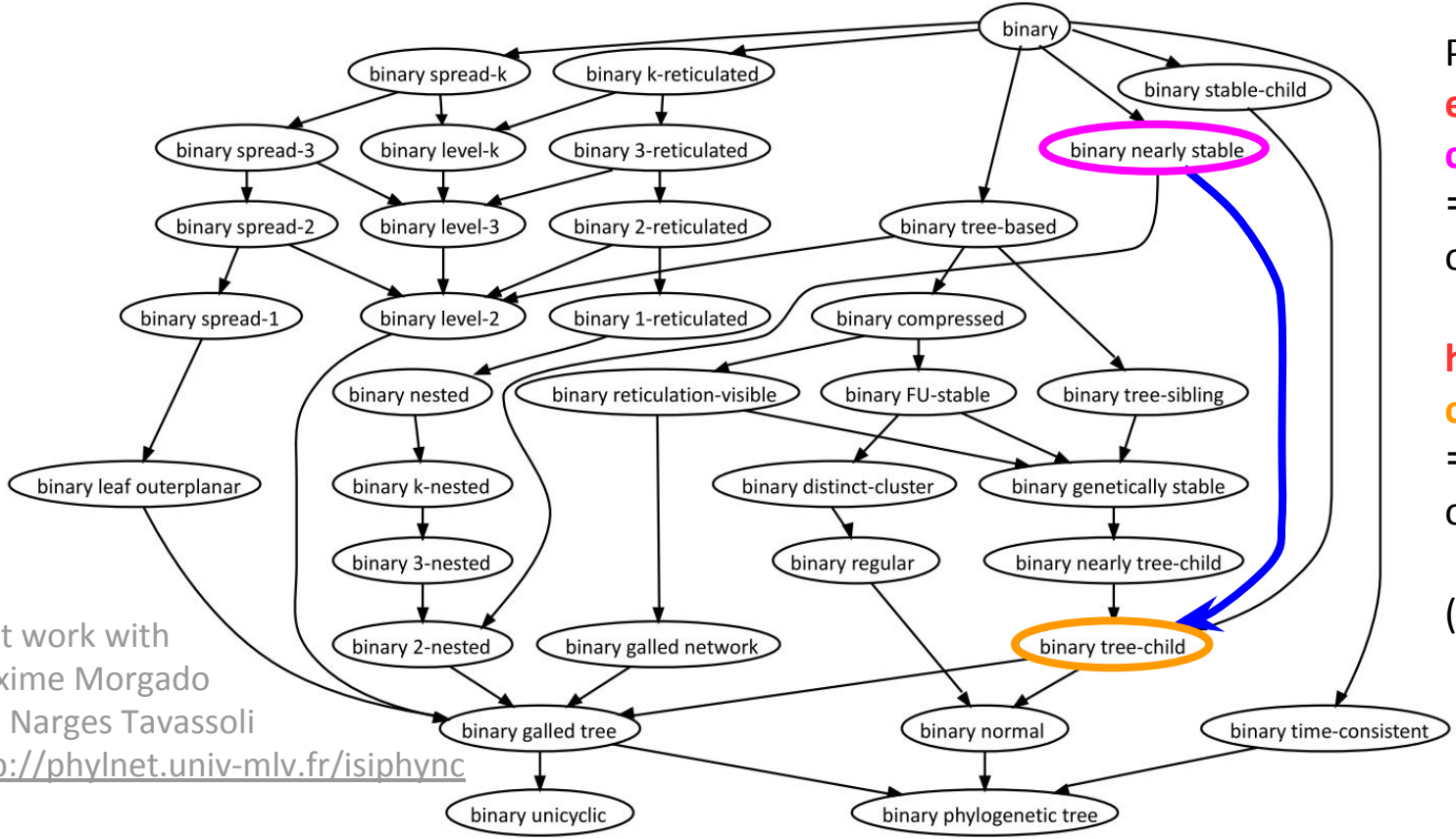


The class of **binary nearly stable networks** contains the class of **binary tree-child networds**:

every **binary tree-child network** is a **binary nearly stable network**

joint work with
Maxime Morgado
and Narges Tavassoli
http://phylnet.univ-mlv.fr/isiphync

Subclasses of phylogenetic networks



Problem
easy to solve on
class A
⇒ **easy to solve**
on **subclass B**

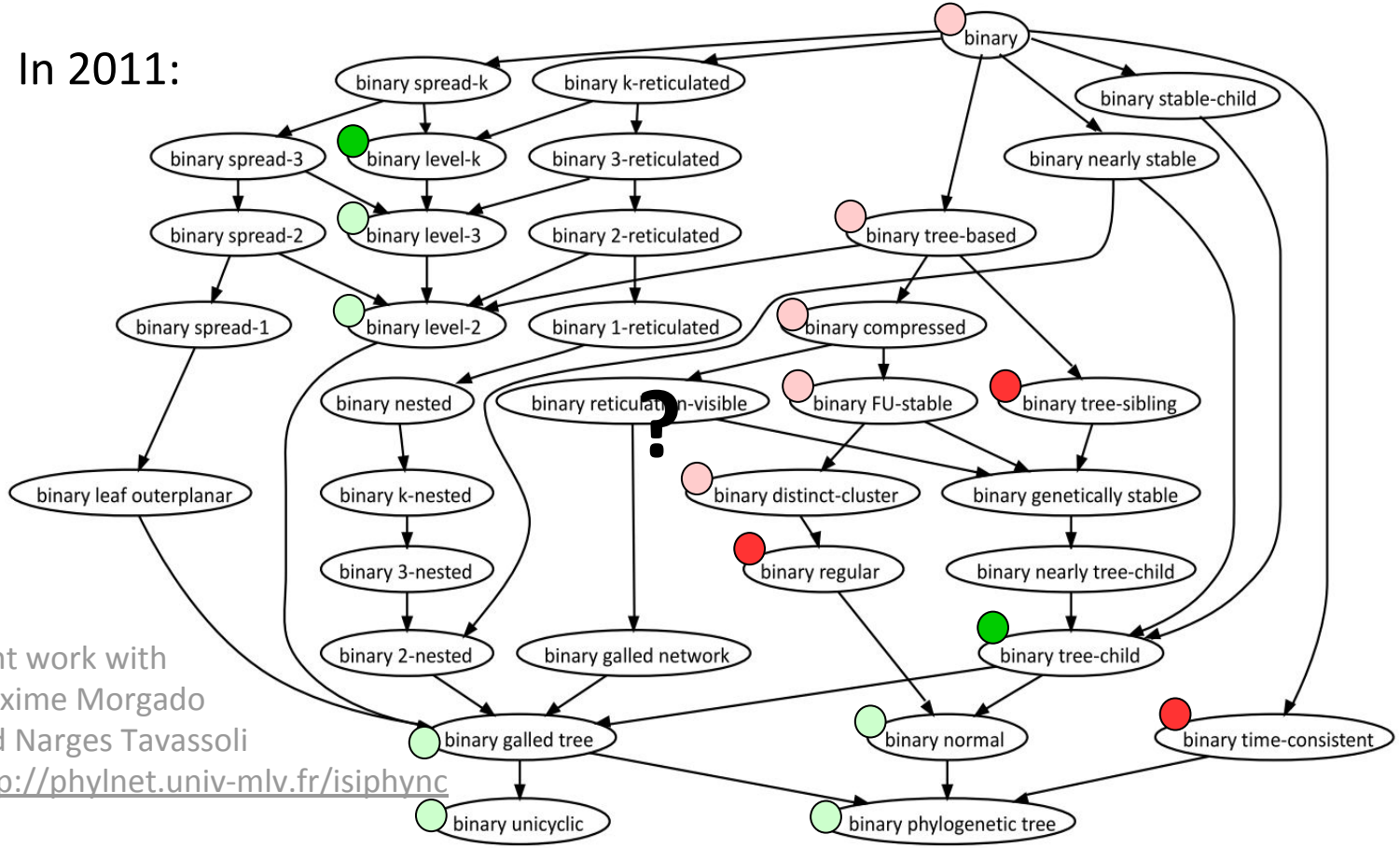
hard to solve on
class B
⇒ **hard to solve**
on **superclass A**

(similar to ISGCI)

joint work with
Maxime Morgado
and Narges Tavassoli
http://phylnet.univ-mlv.fr/isiphync

Understanding tree containment

In 2011:



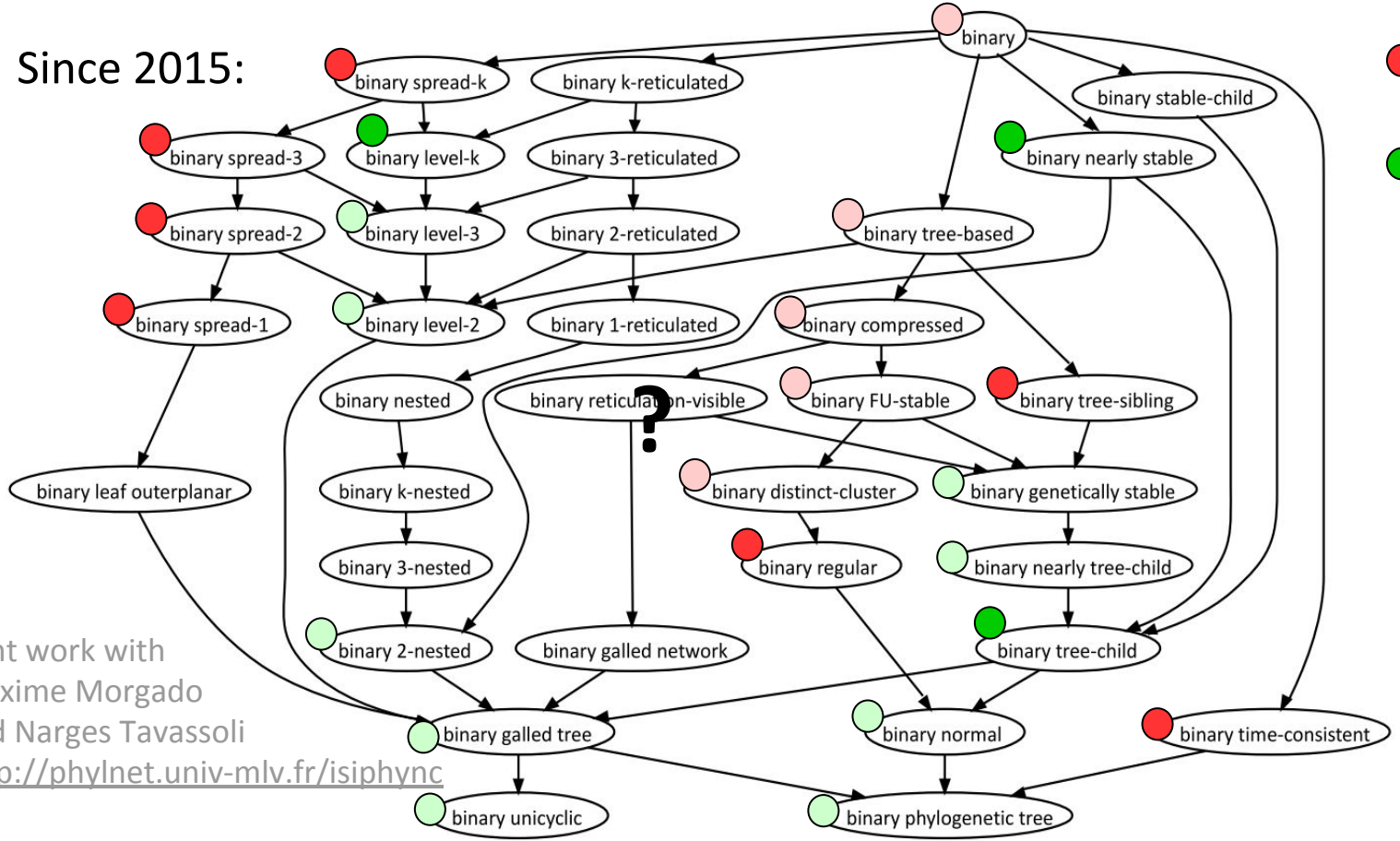
NP-hard

polynomial-time solvable

joint work with
Maxime Morgado
and Narges Tavassoli
http://phylnet.univ-mlv.fr/isiphync

Understanding tree containment

Since 2015:



NP-hard

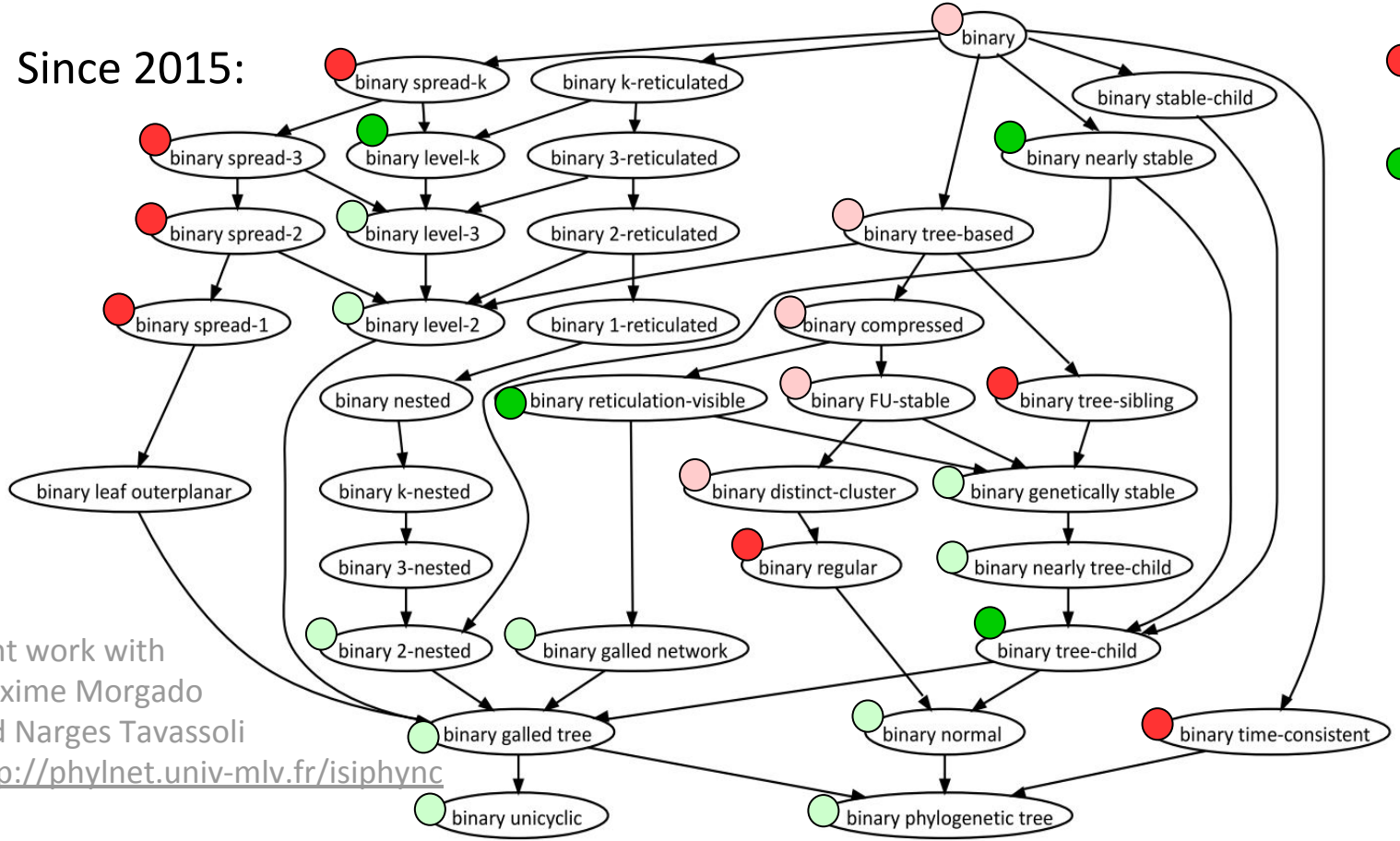
polynomial-time solvable

Results obtained with Andreas Gunawan, Anthony Labarre, Stéphane Vialette & Louxin Zhang

joint work with Maxime Morgado and Narges Tavassoli
http://phylnet.univ-mlv.fr/isiphync

Understanding tree containment

Since 2015:



NP-hard

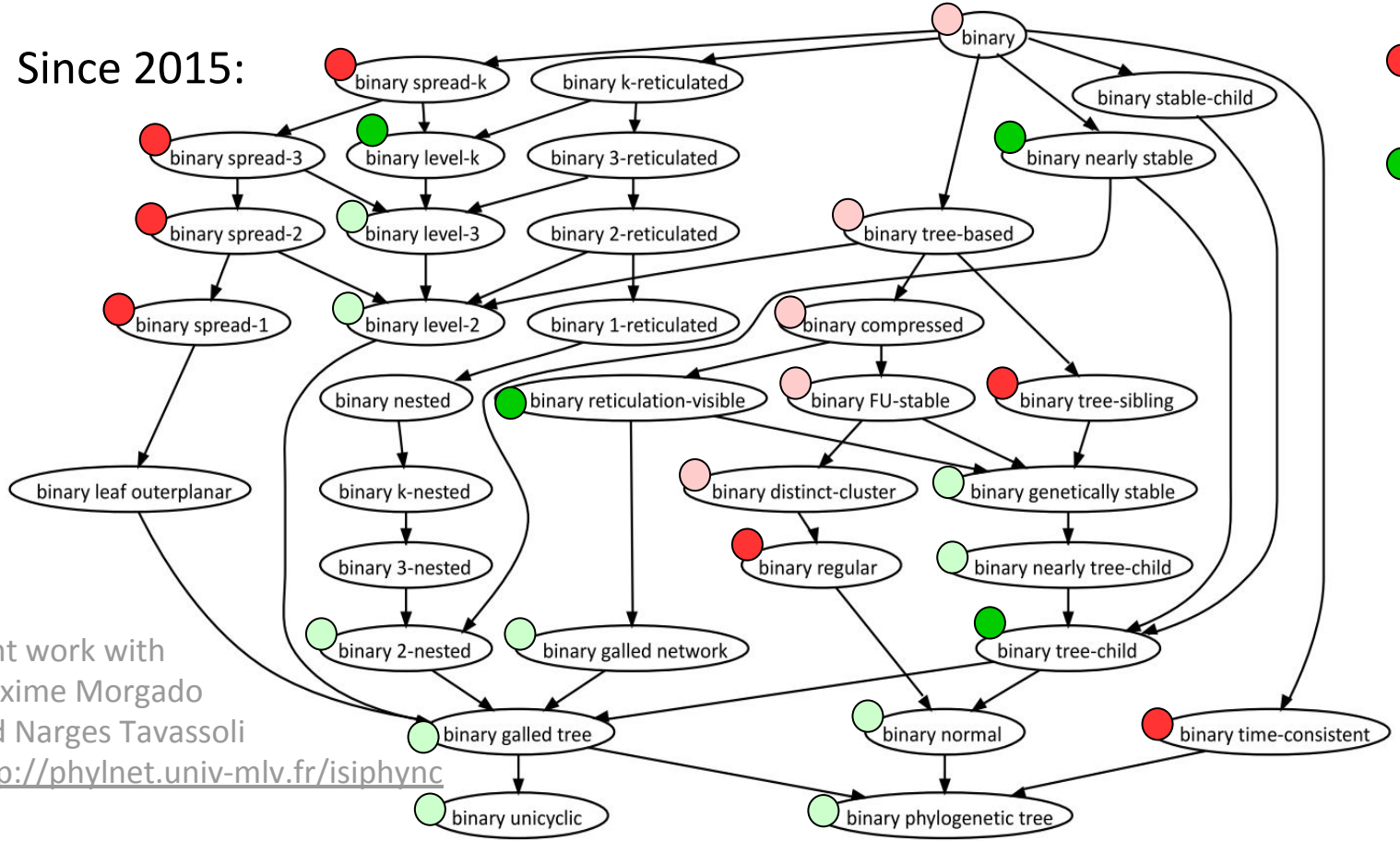
polynomial-time solvable

Polynomial-time algorithm found for binary reticulation-visible networks
Gunawan, Bhaskar DasGupta & Zhang 2016
Weller 2018

joint work with
Maxime Morgado
and Narges Tavassoli
http://phylnet.univ-mlv.fr/isiphync

Understanding tree containment

Since 2015:



NP-hard

polynomial-time solvable

In 2020, first results on the Network Containment Problem for tree-child networks

Janssen & Murakami, AlCoB 2020

joint work with Maxime Morgado and Narges Tavassoli
http://phylnet.univ-mlv.fr/isiphync

KEEP CALM AND KNOW YOUR NETWORK SPACE

Bounding the size of phylogenetic networks

How many nodes can a network on *n* leaves have?

→ unbounded for general networks

→ for nearly-stable networks:

- 26*n*-24

Philippe Gambette, Andreas Gunawan, Anthony Labarre, Stéphane Vialette and Louxin Zhang. Locating a Tree in A Phylogenetic Network in Quadratic Time. *RECOMB 2015*

- 8*n*-7

Andreas Gunawan and Louxin Zhang. *Bounding the Size of a Network Defined By Visibility Property*. arXiv, 2015.

Counting phylogenetic networks

How big is the **search space**?

→ analytic combinatorics techniques to count the number of networks in some subclasses

n	g_{n-1}	r_n	u_{n-1}	ℓ_n
1	0	1	0	1
2	1	3	1	18
3	2	36	6	1 143
4	15	723	135	120 078
5	192	20 280	5 052	17 643 570
6	3 450	730 755	264 270	3 332 111 850
as $n \to \infty$	$c_1 \approx 0.20748$	$c_1 \approx 0.1339$	$c_1 \approx 0.07695$	$c_1 \approx 0.02931$
$x_n \sim c_1 c_2^n n^{n-1}$ with	$c_2 \approx 1.89004$	$c_2 \approx 2.943$	$c_2 \approx 5.4925$	$c_2 \approx 15.4333$
OEIS reference	A328121	A328122	A333005	A333006

Mathilde Bouvel, Philippe Gambette & Marefatollah Mansouri, Counting Phylogenetic Networks of Level 1 and 2, accepted to *JOMB*

Counting phylogenetic networks

How big is the **search space**?

⟶ analytic combinatorics techniques to count the number of networks in some subclasses

$$\phi(z)^n = \sum_{i\geq 0}\sum_{k=0}^{i}\sum_{p=0}^{k}\sum_{q=0}^{p}\sum_{s=0}^{q} \times \binom{n+i-1}{i}\binom{i}{k}\binom{k}{p}\binom{p}{q}\binom{q}{s}\left(\frac{12z}{4(1-z)^4}\right)^{i-k}\left(\frac{-30z^2}{4(1-z)^4}\right)^{k-p}$$

$$\left(\frac{32z^3}{4(1-z)^4}\right)^{p-q}\left(\frac{-16z^4}{4(1-z)^4}\right)^{q-s}\left(\frac{3z^5}{4(1-z)^4}\right)^{s}$$

$$= \sum_{i\geq 0}\sum_{k=0}^{i}\sum_{p=0}^{k}\sum_{q=0}^{p}\sum_{s=0}^{q}\binom{n+i-1}{i}\binom{i}{k}\binom{k}{p}\binom{p}{q}\binom{q}{s}\frac{(3)^i(\frac{-15}{6})^k(\frac{-16}{15})^p(\frac{-1}{2})^q(\frac{-3}{16})^s}{(1-z)^{4i}}$$

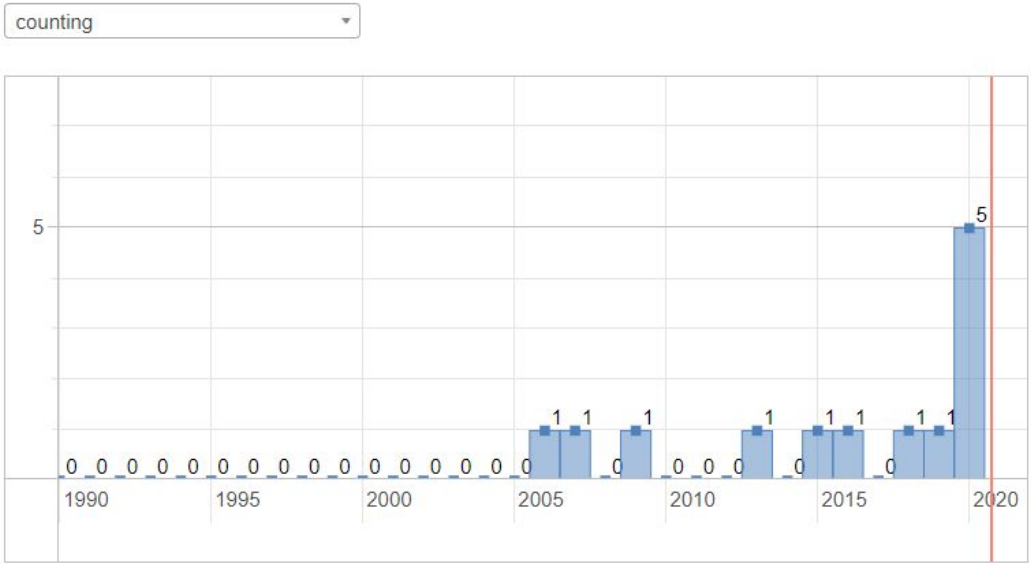
$$\times z^{i+k+p+q+s}.$$

Mathilde Bouvel, Philippe Gambette & Marefatollah Mansouri, Counting Phylogenetic Networks of Level 1 and 2, accepted to *JOMB*

Counting phylogenetic networks

How big is the **search space**?

→ analytic combinatorics techniques to count the number of networks in some subclasses

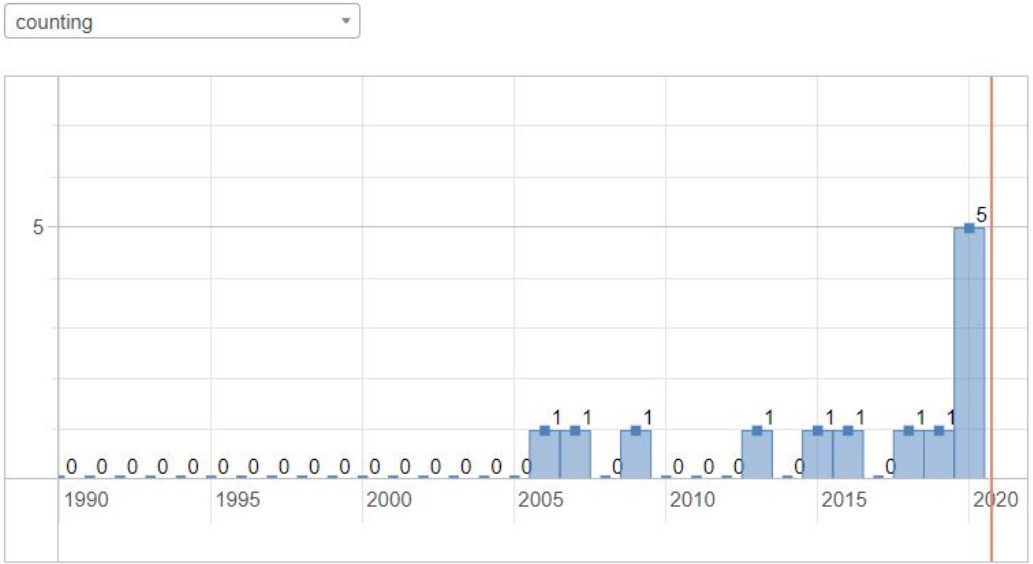


Counting phylogenetic networks

How big is the **search space**?

⟶ analytic combinatorics techniques to count the number of networks in some subclasses

The next step: random generation of phylogenetic networks?



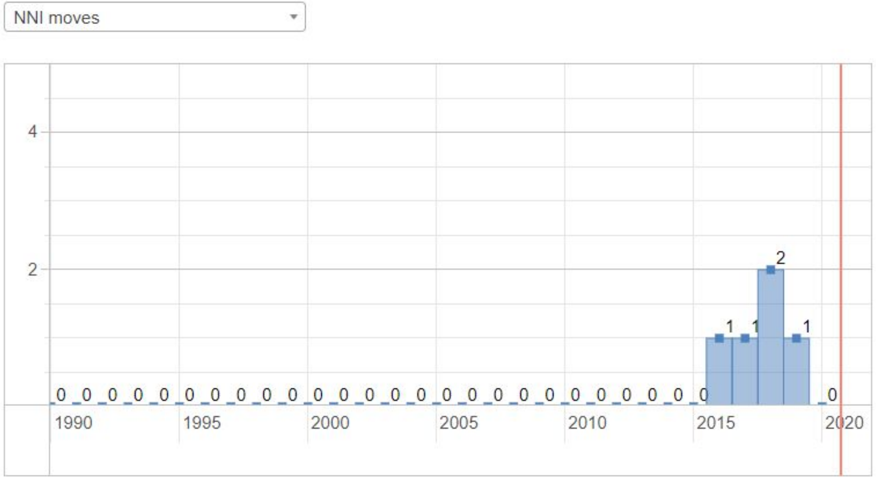
Local moves

How to explore the search space?

→ NNI moves

Katharina Huber, Vincent Moulton and Taoyang Wu. Transforming phylogenetic networks: Moving beyond tree space. *JTB* 404:30-39, 2016.

Philippe Gambette, Leo van Iersel, Mark Jones, Manuel Lafond, Fabio Pardi and Celine Scornavacca. Rearrangement Moves on Rooted Phylogenetic Networks. *PLoS Computational Biology* 13(8): e1005611.1-21, 2017.



Local moves

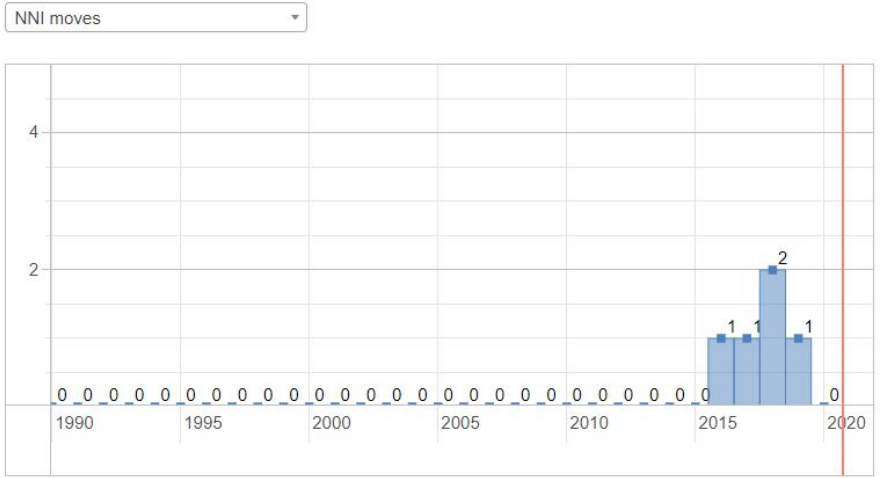
How to explore the search space?

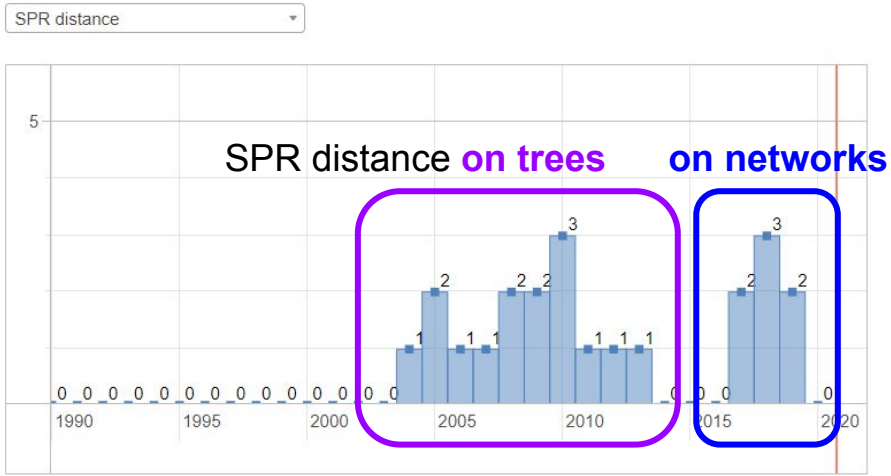
→ NNI moves, SPR moves

Katharina Huber, Vincent Moulton and Taoyang Wu. Transforming phylogenetic networks: Moving beyond tree space. *JTB* 404:30-39, 2016.

Philippe Gambette, Leo van Iersel, Mark Jones, Manuel Lafond, Fabio Pardi and Celine Scornavacca. Rearrangement Moves on Rooted Phylogenetic Networks. *PLoS Computational Biology* 13(8): e1005611.1-21, 2017.

Magnus Bordewich, Simone Linz and Charles Semple. Lost in space? Generalising subtree prune and regraft to spaces of phylogenetic networks. *JTB* 423:1-12, 2017

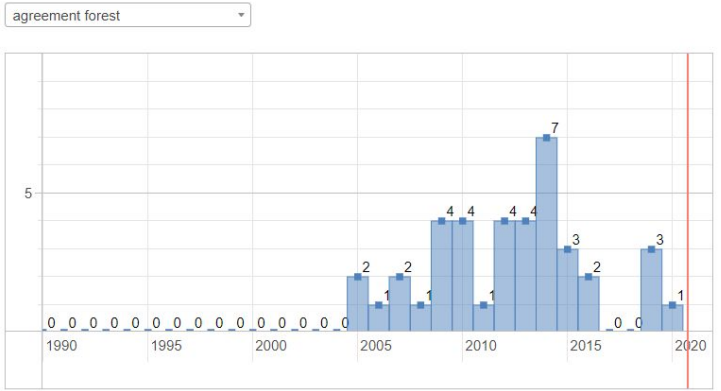




KEEP CALM AND FIND NEW TECHNIQUES

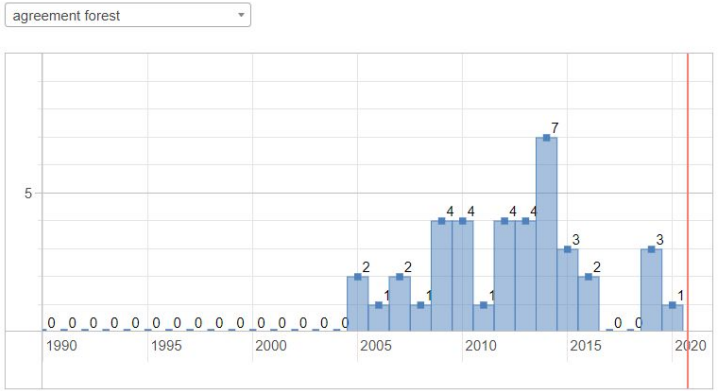
New techniques developed for phylogenetic networks

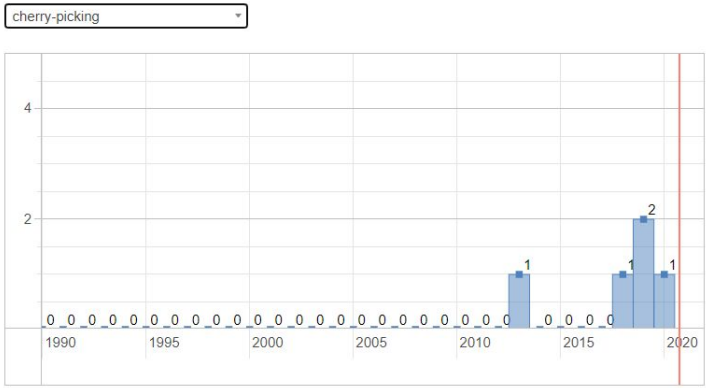
- **agreement forests**: to compute the SPR distance between trees and to solve the hybridization problem between 2 trees



New techniques developed for phylogenetic networks

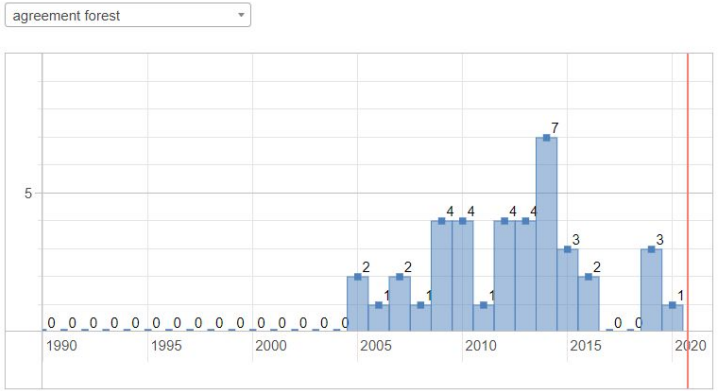
- **agreement forests**: to compute the SPR distance between trees and to solve the hybridization problem between 2 trees
- **cherry picking**: to solve the hybridization problem between > 2 trees

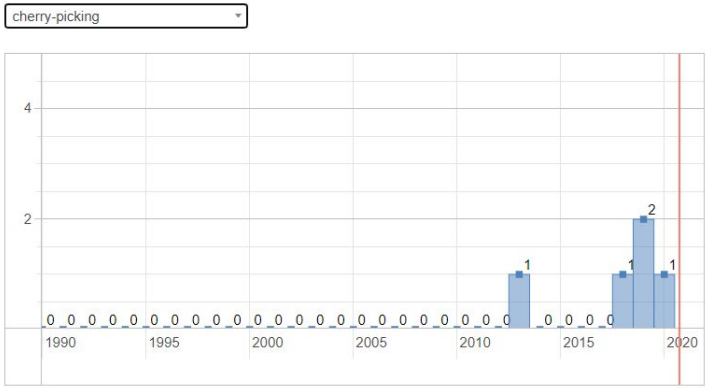




New techniques developed for phylogenetic networks

- **agreement forests**: to compute the SPR distance between trees and to solve the hybridization problem between 2 trees
- **cherry picking**: to solve the hybridization problem between > 2 trees
- **network decompositions**: to solve the tree containment problem on reticulation visible networks

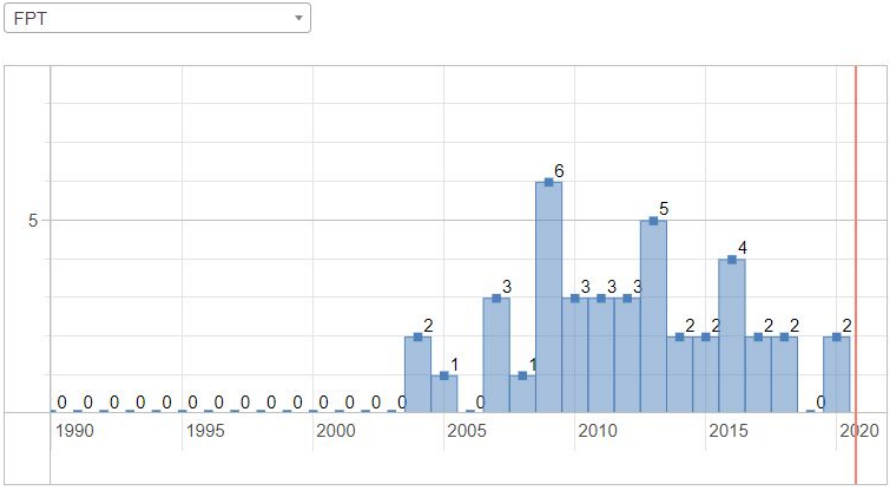




KEEP CALM AND USE POWERFUL TOOLS

Fixed parameter tractability (FPT algorithms)

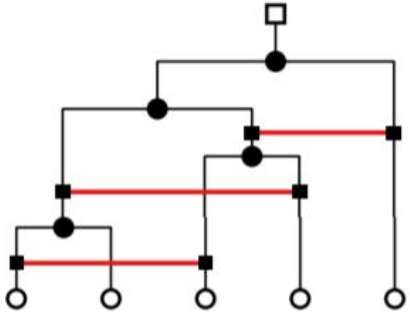
- find an appropriate parameter *k* which is **small**: number of reticulations, level of the network, etc.
- look for an FPT algorithm in *k*: computation time in O(f(*k*)×poly(*n*))
  - computation time may be huge depending on *k*
  - the problem remains tractable when *n* (the number of taxa) increases



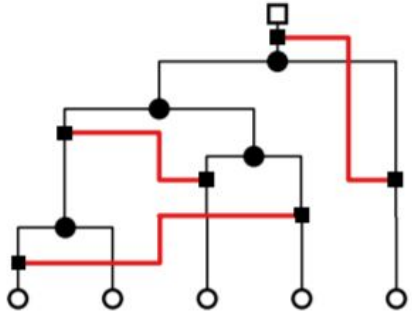
Laurent Bulteau & Mathias Weller, Parameterized Algorithms in Bioinformatics: An Overview, *Algorithms* 12(12):256, 2019

Visualization minimizing edge crossings

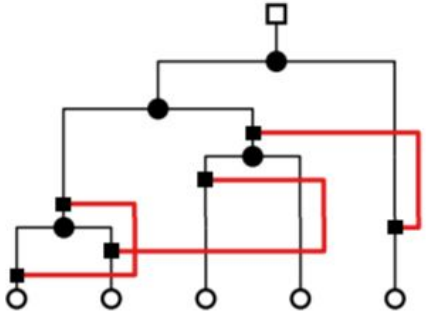
NP-hard
horizontal-style

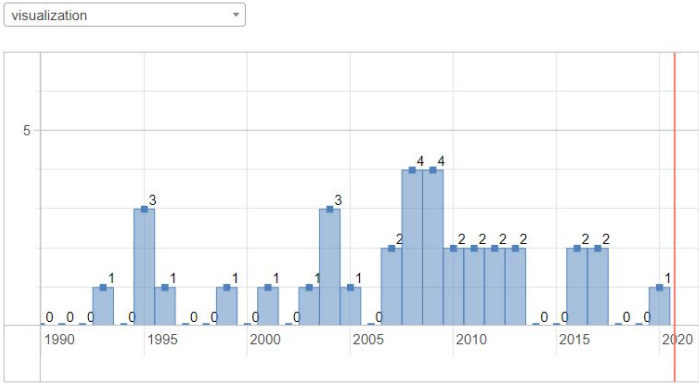


FPT algorithm
snake-style



polynomial-time solvable!
ear-style

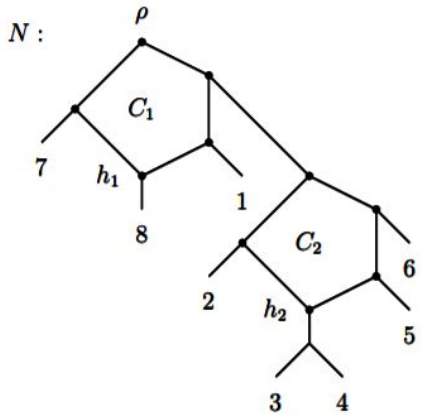




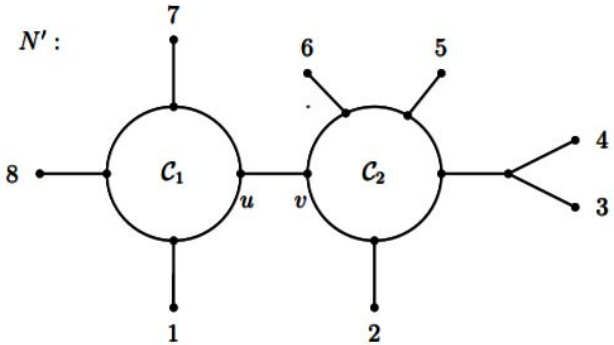
Jonathan Klawitter & Peter Stumpf. Drawing Tree-Based Phylogenetic Networks with Minimum Number of Crossings. arXiv preprint, 2020

Use mathematical properties of abstract networks

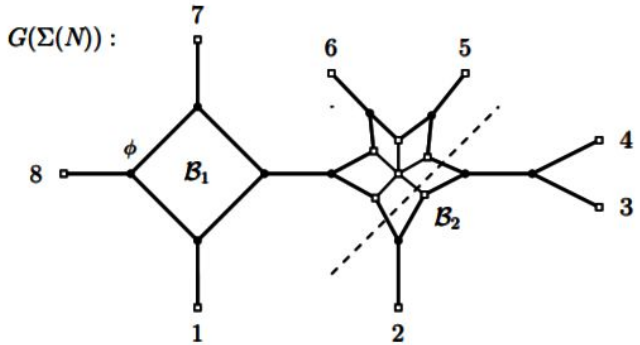
explicit rooted network

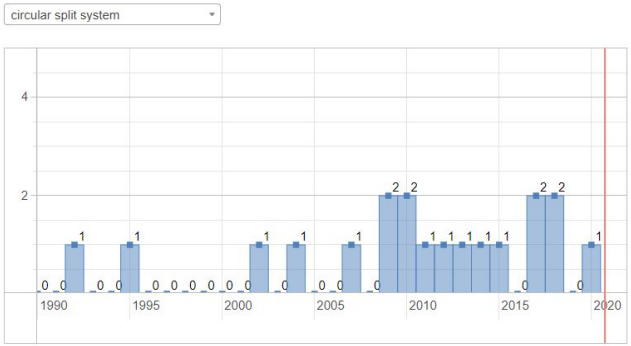


unrooted network



circular split network



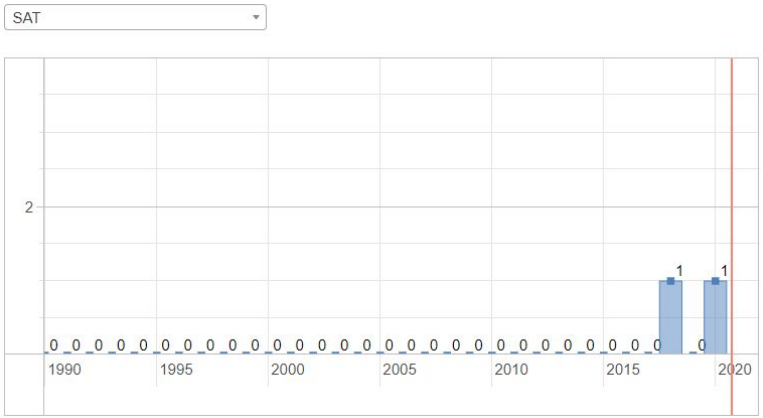


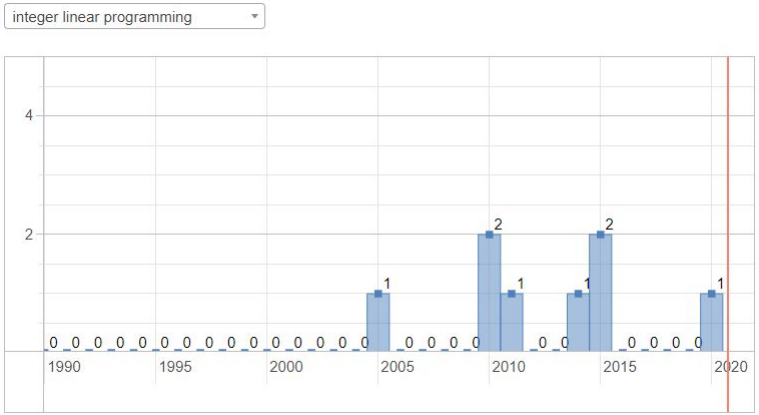
Philippe Gambette, Vincent Berry & Christophe Paul: Quartets and Unrooted Phylogenetic Networks, *JBCB* 10(4):1250004, 2012

Philippe Gambette, Katharina Huber & Guillaume Scholz, Uprooted phylogenetic networks, *BMB*, 79(9):2022-204, 2017

Use solvers

- SAT, ILP (integer linear programming), CSP (constraint satisfaction problem), maximum clique solvers are available





- work in progress with Pierre Bourhis and Marion Tommasi:
  - an ad hoc algorithm is faster most of the time
  - the time taken by the solver does not vary much: more efficient when the ad hoc algorithm takes too long

KEEP CALM AND PUT EVERYTHING TOGETHER

Put everything together

- requires some good engineering work: use multicore processors, parallel or distributed computing, etc.
- requires easy-to-use software:
  - cross-platform software: SplitsTree (1998), Dendroscope (2007), PhyloSketch (2020)
  - web applications: T-REX online (2012)
  - packages or pipeline bricks: R package Phangorn (2011), Julia package PhyloNetworks (2017)

KEEP CALM AND BUILD PHYLOGENETIC NETWORKS

phylnet.info