

Finding Maximum Common Contractions Between Phylogenetic Networks

Bertrand Marchand

Manuel Lafond

Nadia Tahiri

Olivier Tremblay-Savard

Université de Sherbrooke

Université de Sherbrooke

Université de Sherbrooke

University of Manitoba

Finding Maximum Common Contractions Between Phylogenetic Networks

Bertrand Marchand

Université de Sherbrooke

Manuel Lafond

Université de Sherbrooke

Nadia Tahiri

Université de Sherbrooke

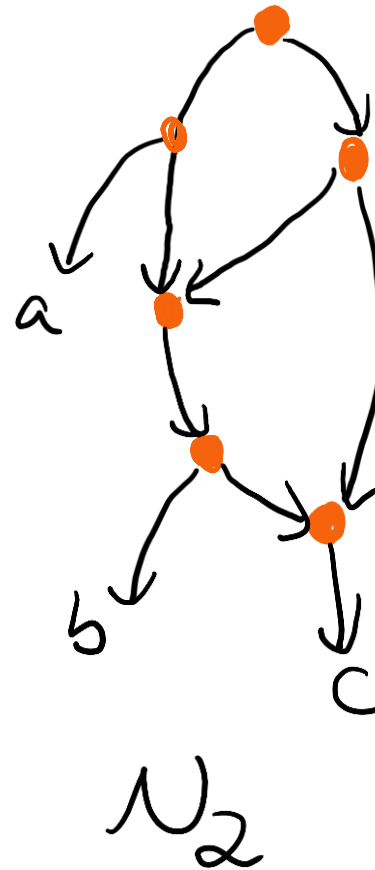
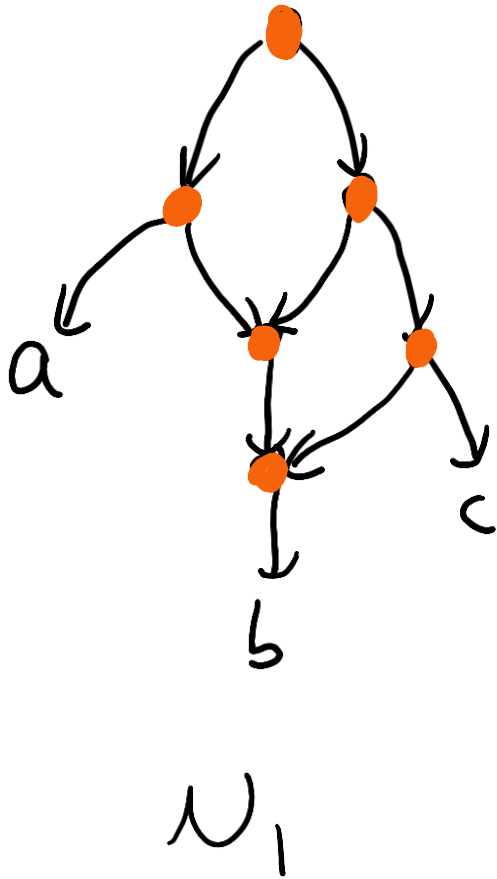
Olivier Tremblay-Savard

University of Manitoba

1. Digressions on metrics and dissimilarity measures
2. Contraction measures: basic properties
3. Algorithmic aspects of contraction measures

Metrics and dissimilarity measures on networks:

- Are these networks the same?
- By how much how they different?
- What do they have in common?



Why metrics and dissimilarity measures

- Compare inferred networks against gold standard
(e.g., PhyNEST used hardwired clusters vs simulated networks)
- Extract common structure.
- Cluster similar networks.

RESEARCH ARTICLE | ANTHROPOLOGY | 



Phylogenetic network analysis of SARS-CoV-2 genomes

Peter Forster , Lucy Forster, Colin Renfrew , and Michael Forster  [Authors Info & Affiliations](#)

Contributed by Colin Renfrew, March 30, 2020 (sent for review March 17, 2020; reviewed by Toomas Kivisild and Carol Stocking)

April 8, 2020 | 117 (17) 9241-9243 | <https://doi.org/10.1073/pnas.2004999117>

Network analysis of SARS-CoV-2

Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary

[Santiago J. Sánchez-Pacheco](#)  ✉, [Sungsik Kong](#) ✉, [Paola Pulido-Santacruz](#) , , and [Laura Kubatko](#) [Authors Info &](#)

[Affiliations](#)

May 7, 2020 | 117 (23) 12518-12519 | <https://doi.org/10.1073/pnas.2007062117>

network analysis of SARS-CoV-2

Median-joining network

Reply to Sánchez-Pacheco et al., Chookajorn, and Mavian et al.: Explaining phylogenetic network analysis of SARS-CoV-2 genomes

Peter Forster ✉, Lucy Forster, Colin Renfrew, and Michael Forster  [Authors Info & Affiliations](#)

May 21, 2020 | 117 (23) 12524-12525 | <https://doi.org/10.1073/pnas.2007433117>

ARTICLE | ANTHROPOLOGY | 

Network analysis of SARS-CoV-2

LETTER | ANTHROPOLOGY | 

Median-joining network

LETTER | ANTHROPOLOGY | 

Sánchez-Pacheco et al., Chookajorn,
phylogenetic

LETTER | ANTHROPOLOGY | 

Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable

Carla Mavian  ✉, Sergei Kosakovsky Pond, Simone Marini,  ⁺³², and Marco Salemi ✉ [Authors Info & Affiliations](#)

May 7, 2020 | 117 (23) 12522-12523 | <https://doi.org/10.1073/pnas.2007295117>

The ideal measure should be:

The ideal measure should be:

- easy to interpret
- easy to compute
- > 0 on different networks
- able to provide common structures
- able to compare any two networks
- be normalizable
- ...

The ideal measure should be:

- easy to interpret
- easy to compute
- > 0 on different networks
- able to provide common structures
- able to compare any two networks
- be normalizable
- ...

Two classes of measures

- **structural**: compares the set of X things shared by the networks
- **operational**: min # of operations to transform one network into the other

Structural measures	Description
Displayed trees	Symmetric diff. between sets of displayed trees
Softwired clusters	Symmetric diff. between sets of softwired clusters
Hardwired clusters	Symmetric diff. between sets of descending leaves
(extended) μ -representation	Diff. between vectors of # node-to-leaf paths
Tri-partition distance	Compares sets of strict vs non-strict descendants
Nodal distance	Compares leaf-to-leaf distances
Triplet distance	Compares subnetworks on three leaves
Operational measures	
rNNI / rSPR / rTBR	Min # of (rooted) NNI / SPR / TBR to turn N1 into N2
Cherry-picking distance	Min # of cherry-picking operations to turn N1 into N2 * or, min # ops on N1 and N2 to reach same network

Structural	>0 if ≠ ?	Compute time	Compare any networks?	Common structure?
Displayed trees	NO	$O(2^r n^c)$	YES	kind of?
Softwired clusters	NO	$O(2^r n^c)$	YES	kind of?
Hardwired clusters	NO	$O(n^2)$	YES	kind of?
(extended) μ -representation	NO *	$O(n^2 \log n)$	YES	NO
Tri-partition distance	NO **	$O(n^4)$	YES	NO
Nodal distance	NO **	$O(n^3)$	YES	NO
Triplet distance	NO **	$O(n^5)$	YES	kind of?
Operational				
rNNI	YES	NP-hard	NO (# retics***)	YES
rSPR / rTBR	YES	$O(2.4^d n)$ Linear kernel	NO (# retics)	YES
Cherry-picking	YES	$O(3^r n^c)$	NO (orchards)	YES

* yes on orchards ** yes on time-consistent tree-child *** add triangle creation move

Structural	References
Displayed trees	[Huson, Rupp, Scornavacca, 2010]
Softwired clusters	-----
Hardwired clusters	[Cardona, Llabrés, Rosselló, Valiente, 2008a]
(extended) μ -distance	[Cardona, Pons, Ribas, Martinez Coronado, 2024]
Tri-partition distance	[Cardona, Llabrés, Rosselló, Valiente, 2008a]
Nodal distance	[Cardona, Llabrés, Rosselló, Valiente, 2008b]
Triplet distance	-----
Operational	
rNNI	[Gambette, Van Iersel, Jones, Lafond, Pardi, Scornavacca, 2017] [Janssen, Jones, L Erdős, van Iersel, Scornavacca, 2018]
rSPR / rTBR	[Bordewich, Linz, Semple, 2017] [Kelk, Linz, Meuwese, 2023]
Cherry-picking	[Landry, Teodocio, L, Tremblay-Savard, 2022-2024]

The ideal measure should be:

- easy to interpret
- easy to compute
- > 0 on different networks
- able to provide common structures
- able to compare any two networks
- be normalizable
- ...

Unrealistic?

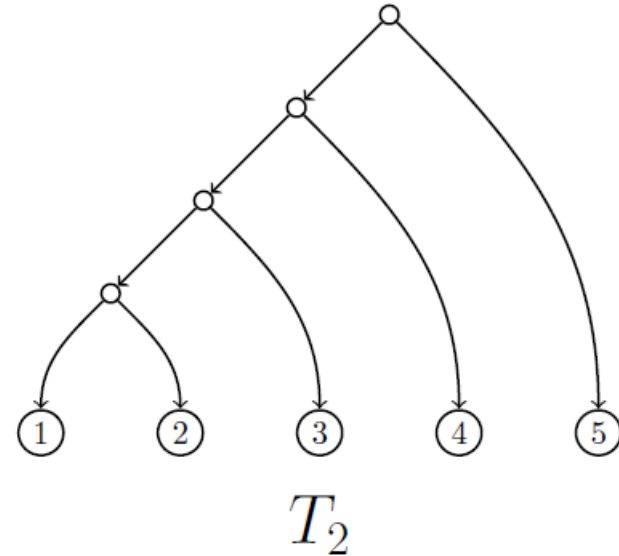
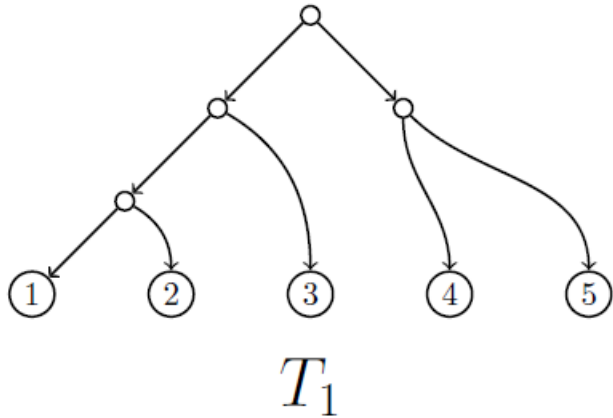
The ideal measure should be:

- easy to interpret
- easy to compute
- > 0 on different networks
- able to provide common structures
- able to compare any two networks
- be normalizable
- ...

Unrealistic?

- seems feasible on trees: Robinson-Foulds metric checks all these boxes and is widely used in practice.

Robinson-Foulds on trees

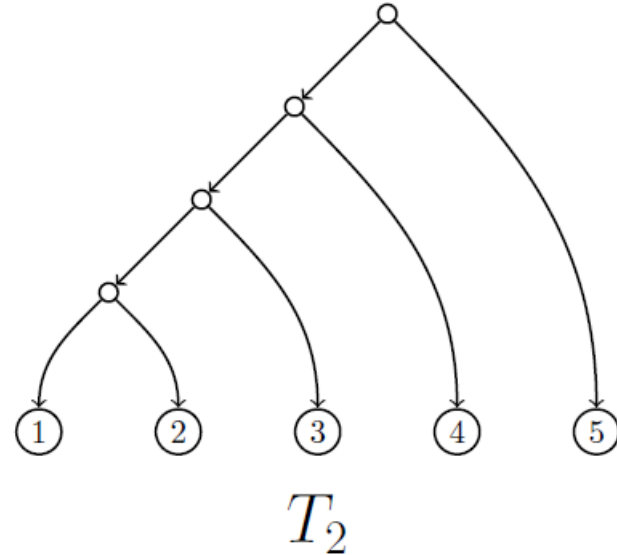
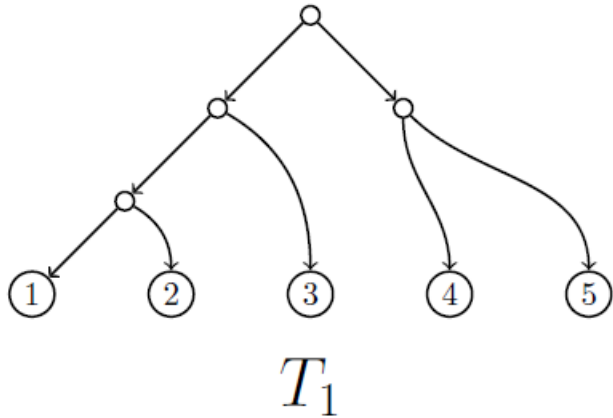


RF: size of symmetric difference of the set of clades

$clades(T_1)$: $\{1,2\}$ $\{1,2,3\}$ $\{4,5\}$

$clades(T_2)$: $\{1,2\}$ $\{1,2,3\}$ $\{1,2,3,4\}$ $RF = 2$

Robinson-Foulds on trees



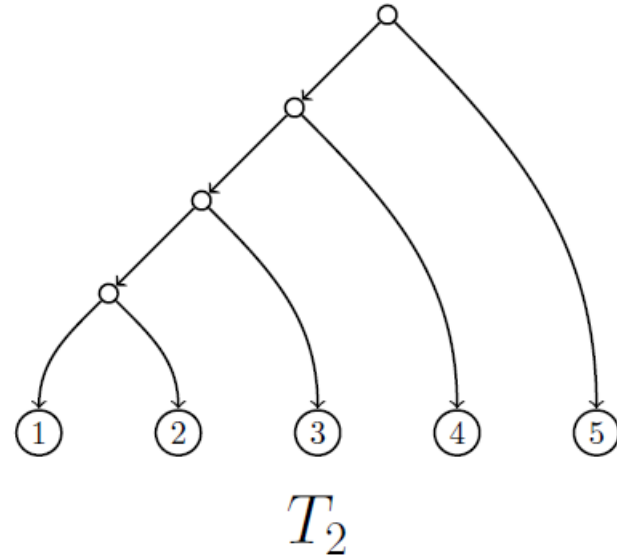
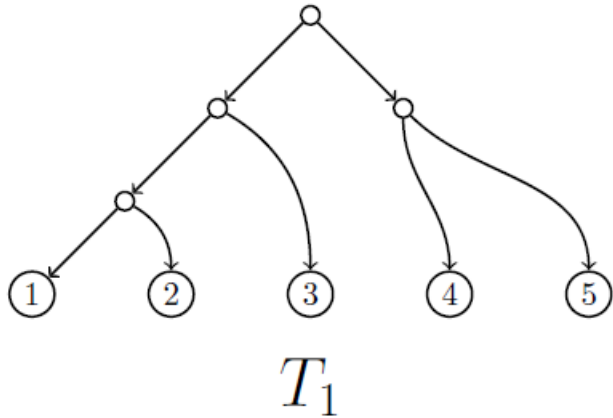
RF: size of symmetric difference of the set of clades

Many structural measures generalize RF (equal to RF on trees)

- Hardwired/softwired clusters, μ -distance, tri-partitions

How about operational generalizations?

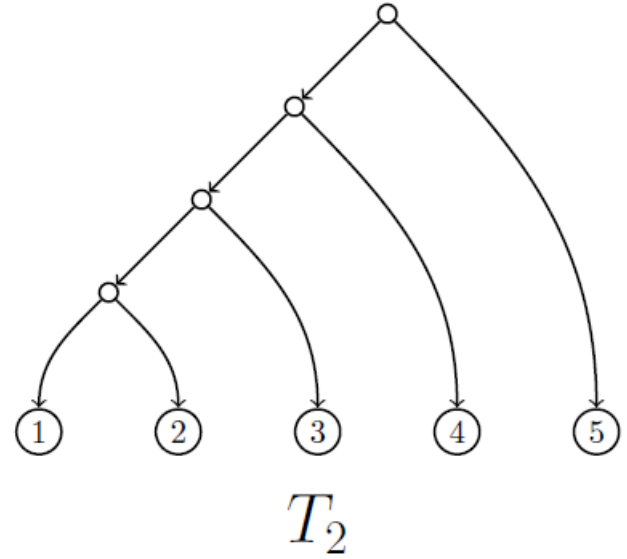
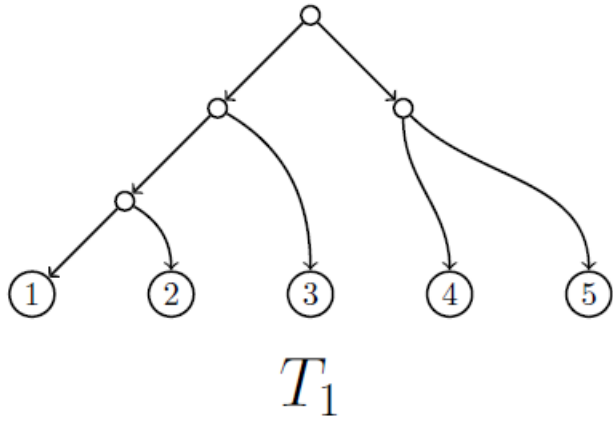
Robinson-Foulds on trees



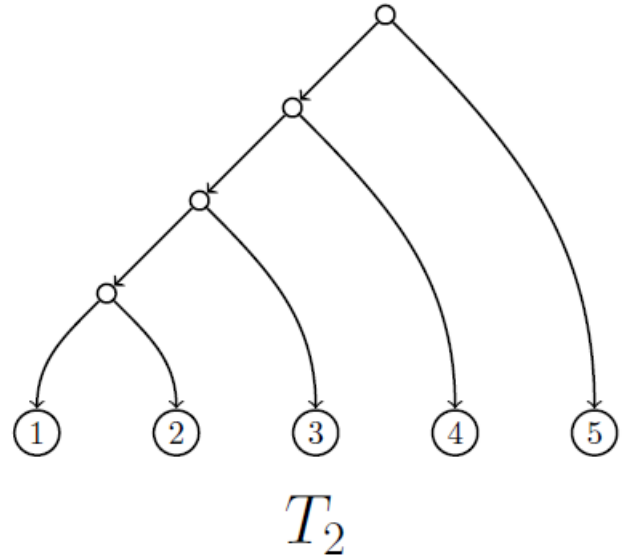
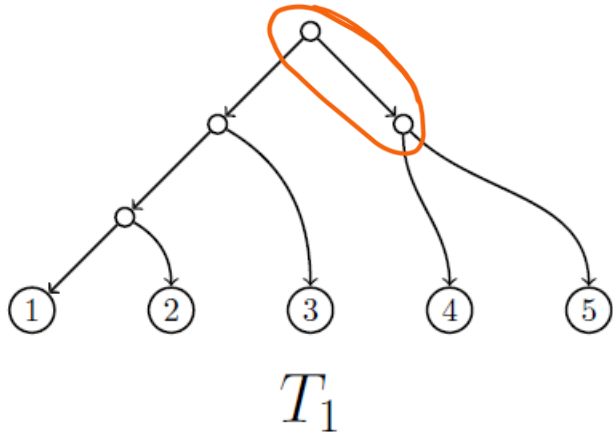
Original definition from [RF81]

Minimum # of edge contractions and expansions required to transform T_1 into T_2 .

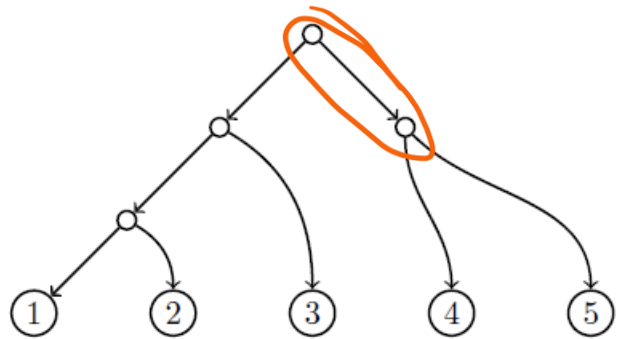
Robinson-Foulds on trees



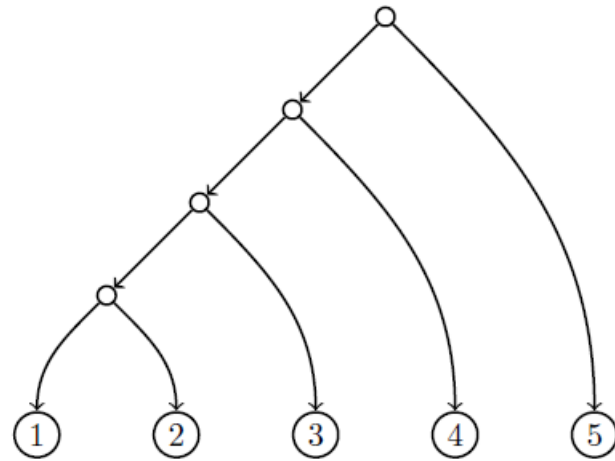
Robinson-Foulds on trees



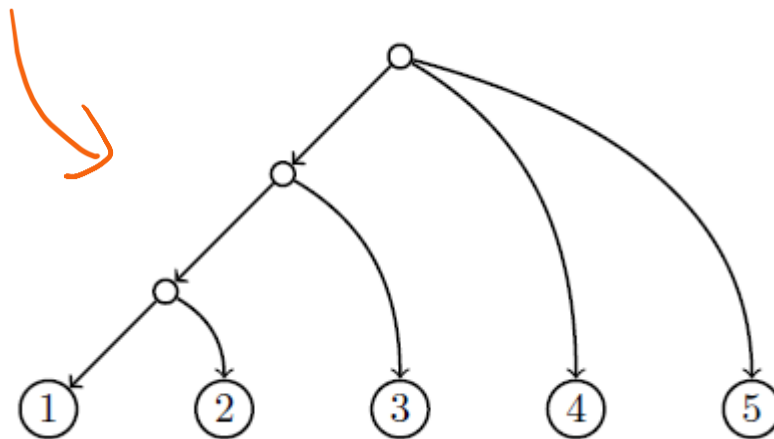
Robinson-Foulds on trees



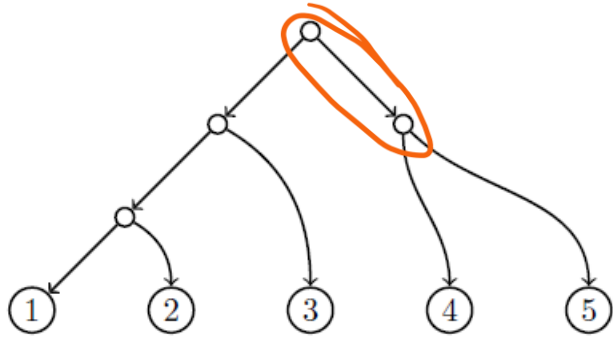
T_1



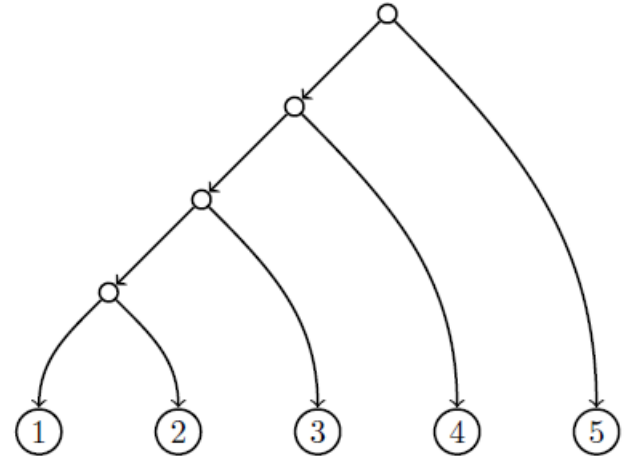
T_2



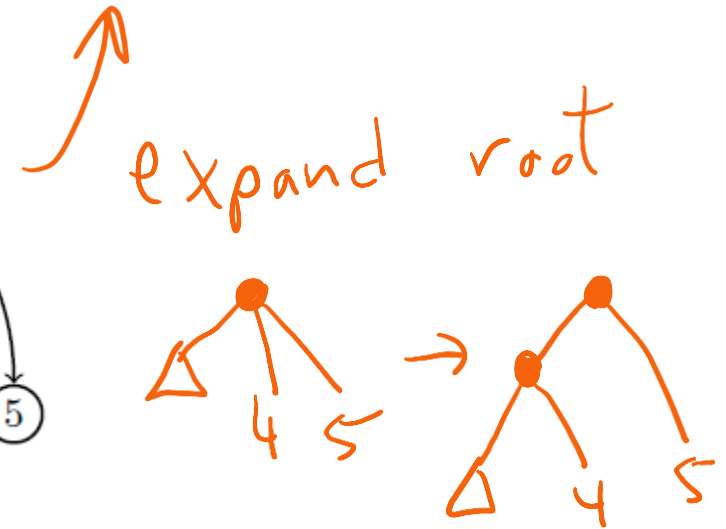
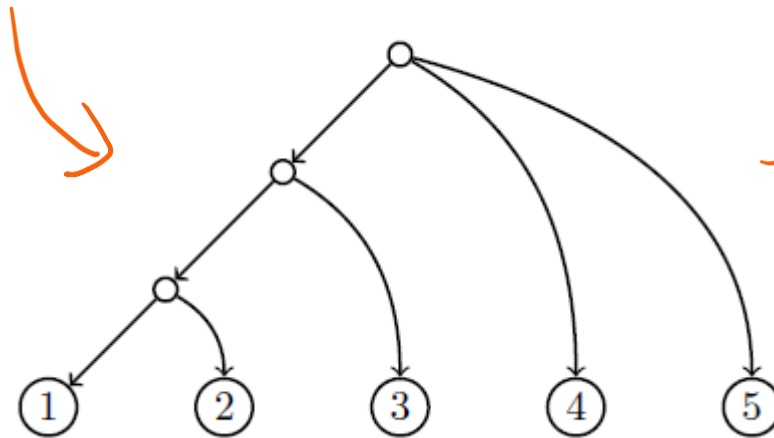
Robinson-Foulds on trees



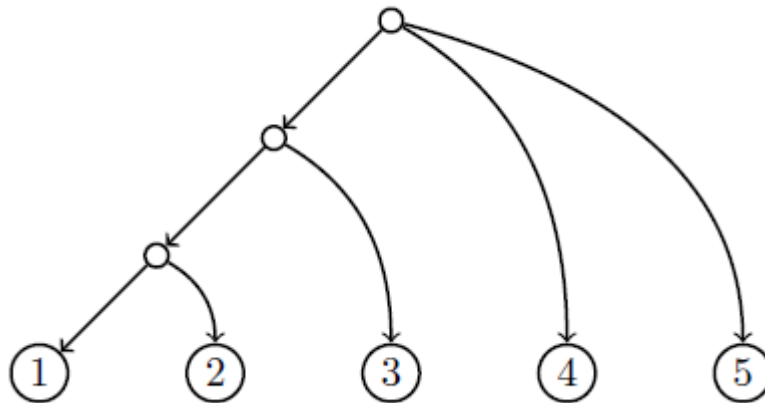
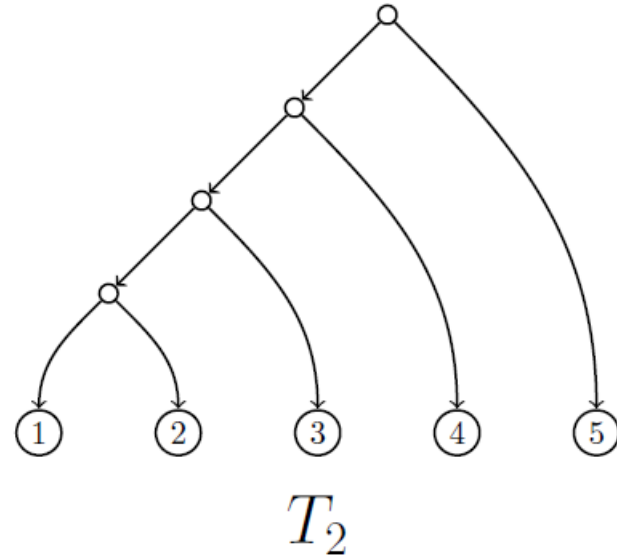
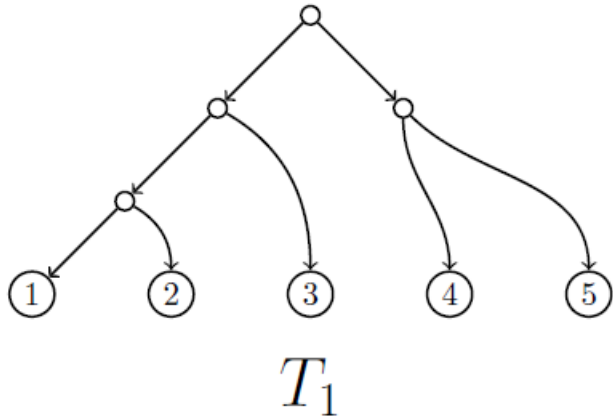
T_1



T_2



Robinson-Foulds on trees



BONUS

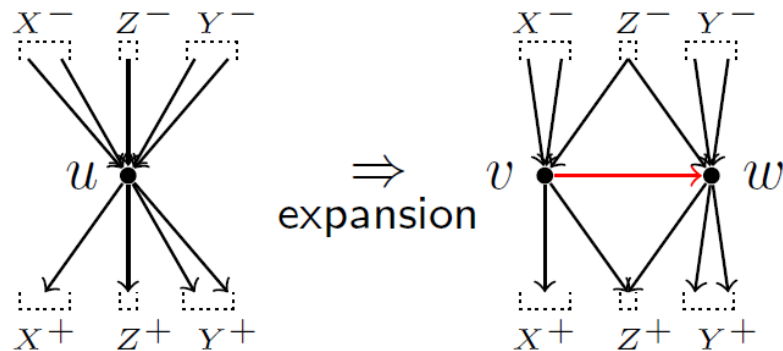
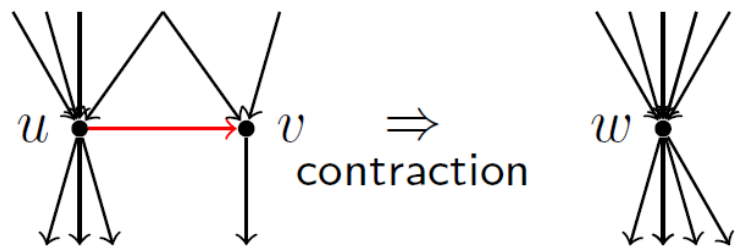
If we contract different edges in either T_1 or T_2 , we get the same tree: a common contraction

Operational	>0 if ≠ ?	Compute time	Common structure?	Compare any networks?
rNNI	YES	?	?	NO (same # retics)
SPR / TBR	YES	$O(2.4^d n)$ Linear kernel	YES	NO (same # retics)
Cherry-picking	YES	$O(3^r n^c)$	YES	NO (orchards)

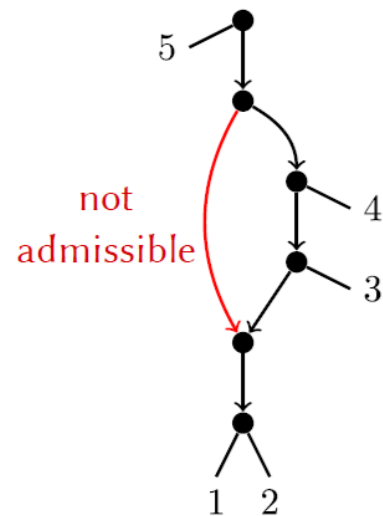
Operational	>0 if ≠ ?	Compute time	Common structure?	Compare any networks?
rNNI	YES	?	?	NO (same # retics)
SPR / TBR	YES	$O(2.4^d n)$ Linear kernel	YES	NO (same # retics)
Cherry-picking	YES	$O(3^r n^c)$	YES	NO (orchards)
Contraction measures	YES		YES	YES

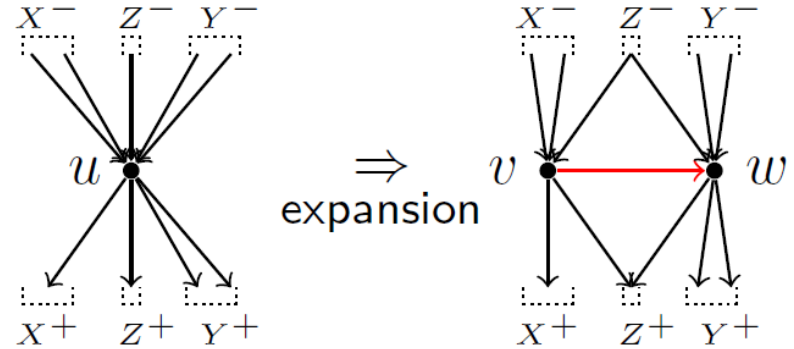
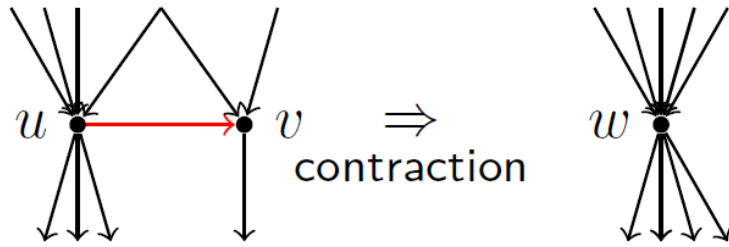
Operational	>0 if ≠ ?	Compute time	Common structure?	Compare any networks?
rNNI	YES	?	?	NO (same # retics)
SPR / TBR	YES	$O(2.4^d n)$ Linear kernel	YES	NO (same # retics)
Cherry-picking	YES	$O(3^r n^c)$	YES	NO (orchards)
Contraction measures	YES	I HAVE NO IDEA HOW TO COMPUTE THIS!	YES	YES

Contraction-based measures and basic properties



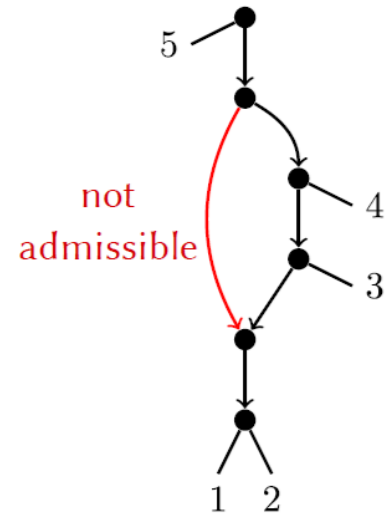
A contraction is **admissible** if it does not create a directed cycle.



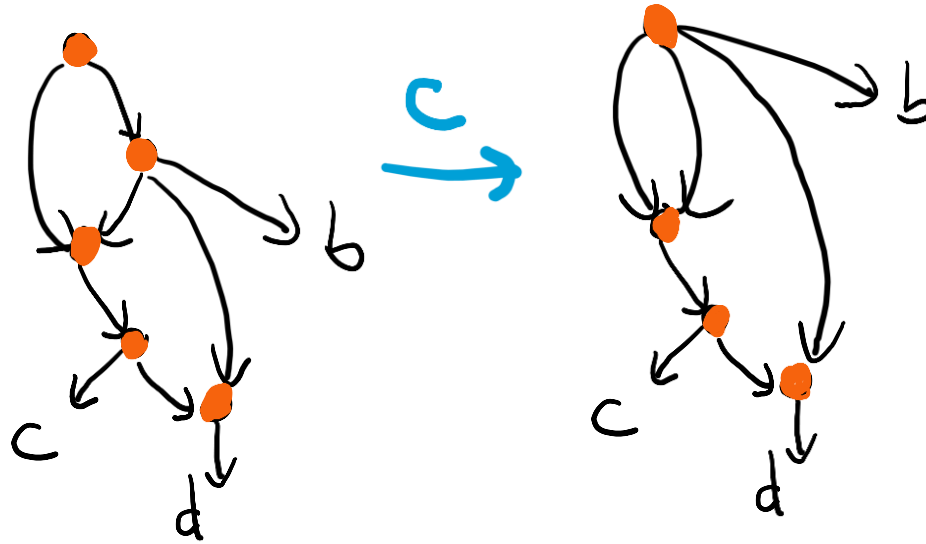


A contraction is **admissible** if it does not create a directed cycle.

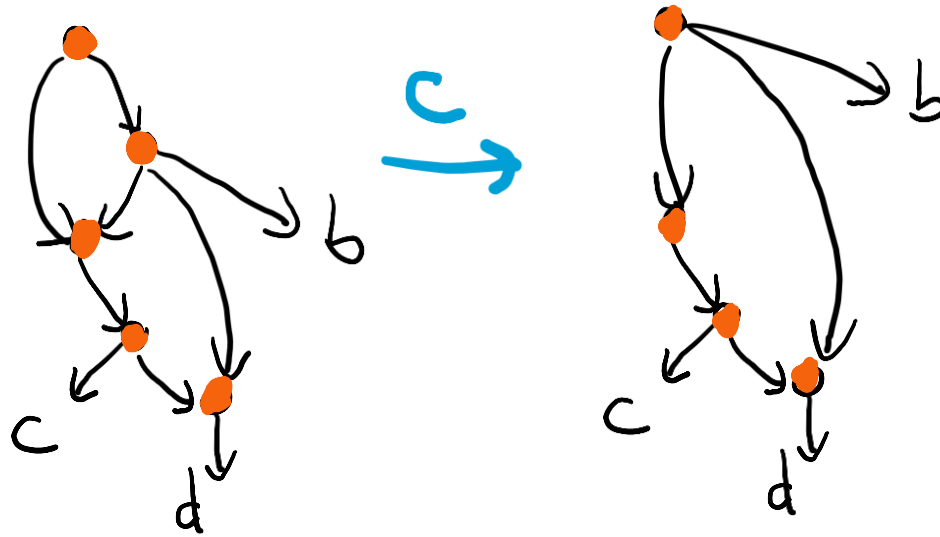
- Also, set of leaves must remain the same
- Also, if parallel edges are created, keep only one.
- Also, we allow degree 2 vertices (indeg = outdeg = 1)



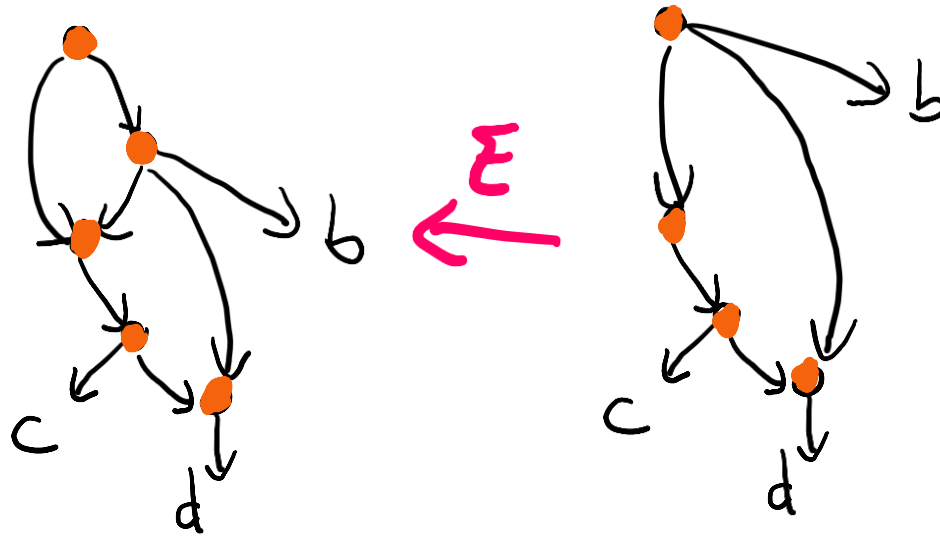
- Contractions can destroy cycles



- Contractions can destroy cycles

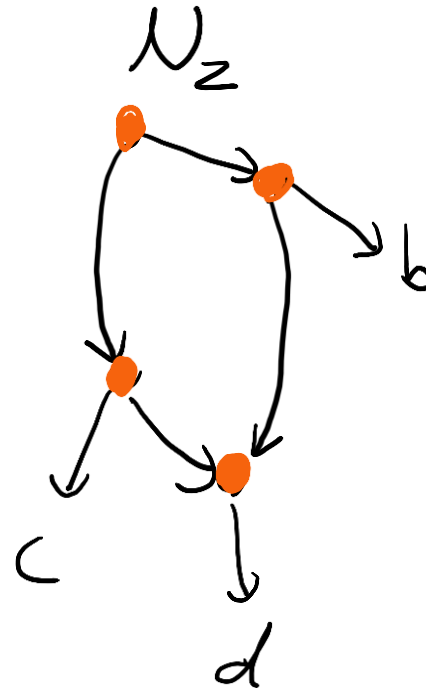
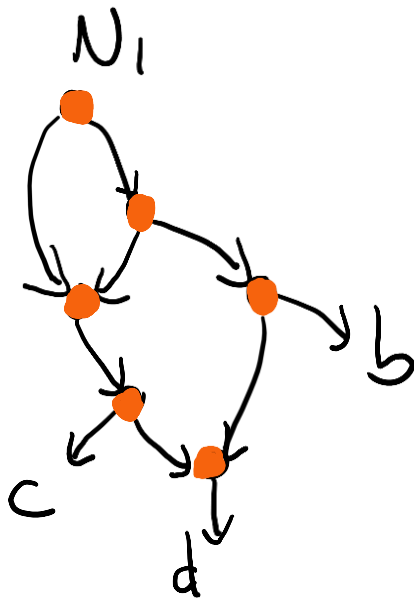


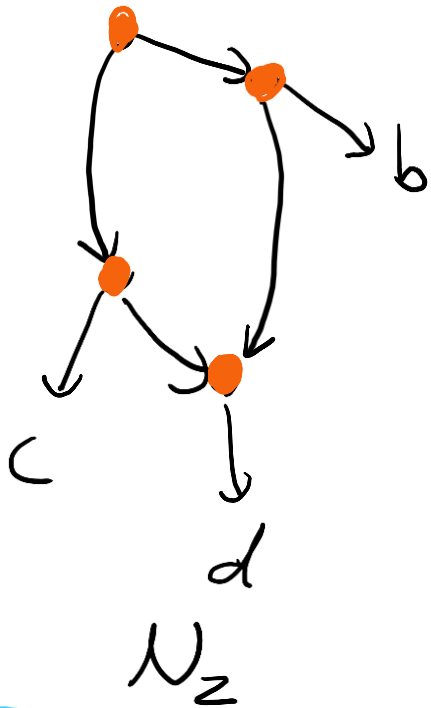
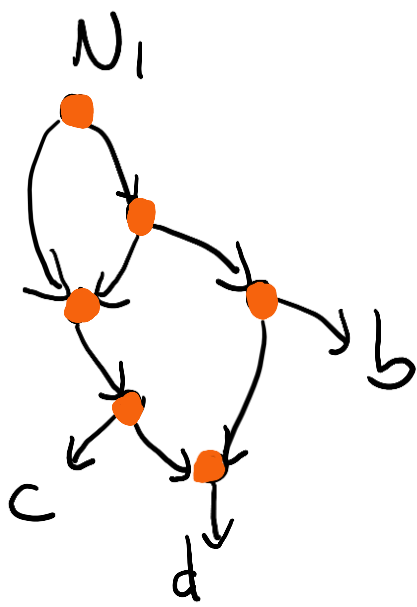
- Expansions can create cycles

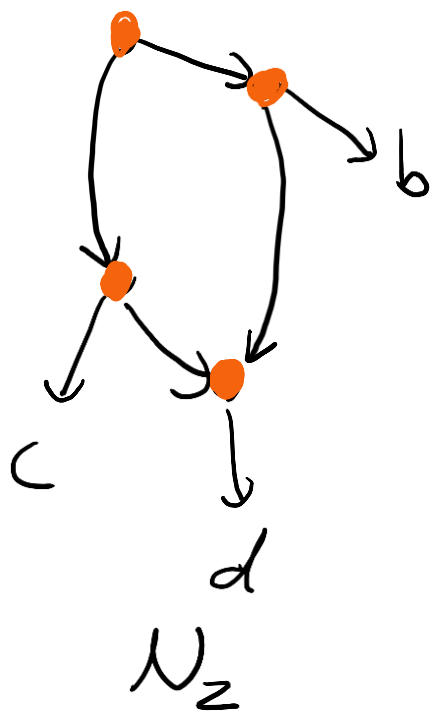
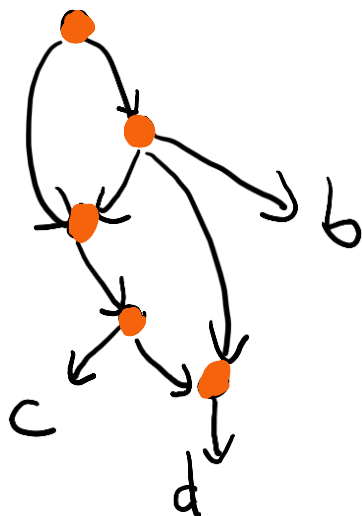
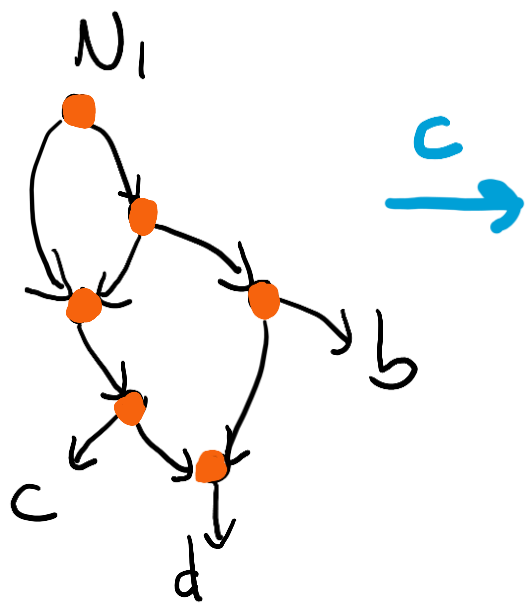


Contraction distance d_{CE}

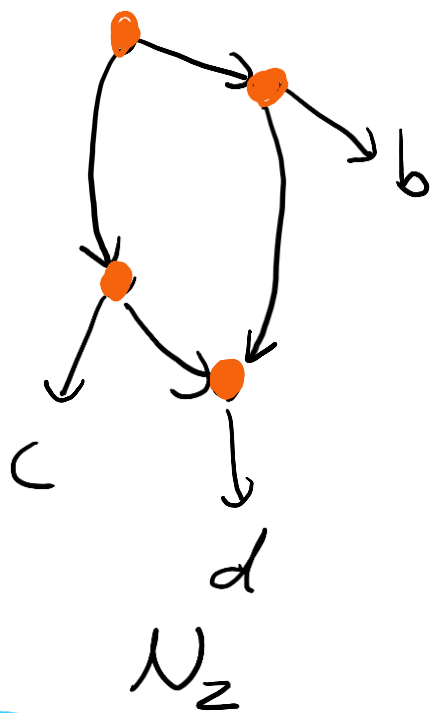
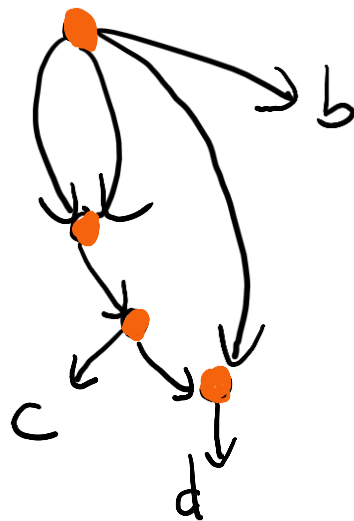
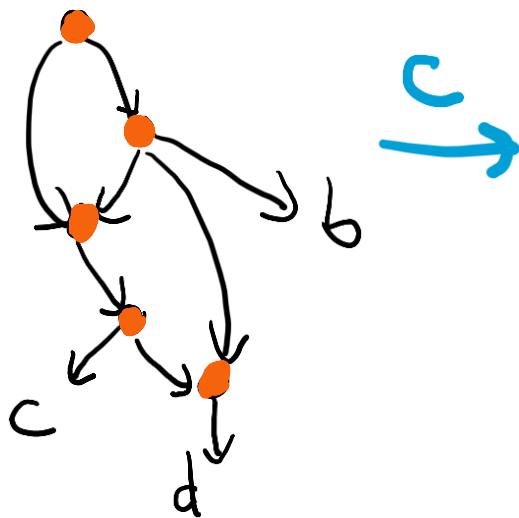
$d_{CE}(N_1, N_2) = \min \#$ of admissible contractions/expansions to transform N_1 into a network isomorphic to N_2 .

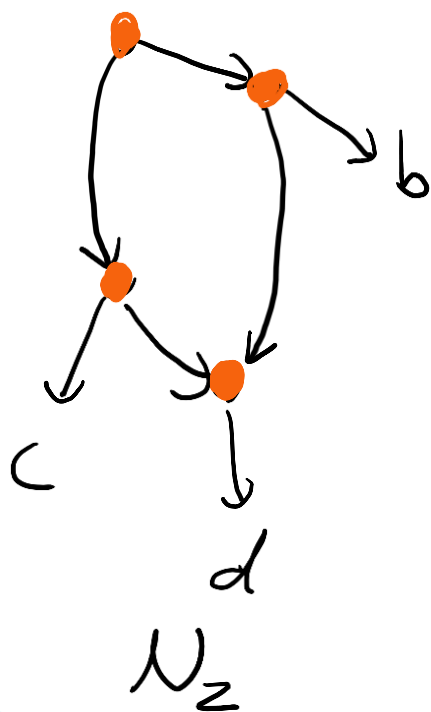
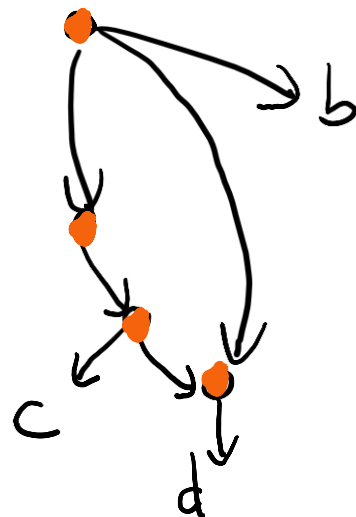
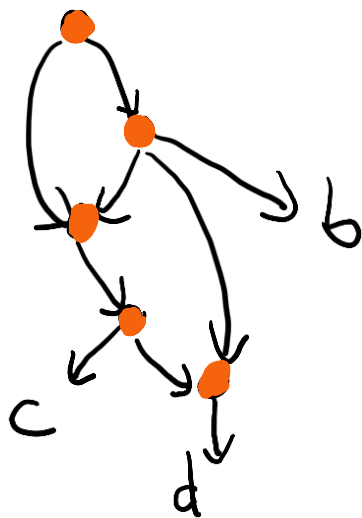
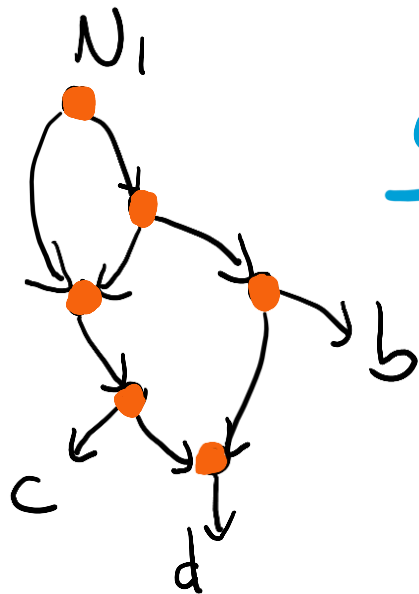


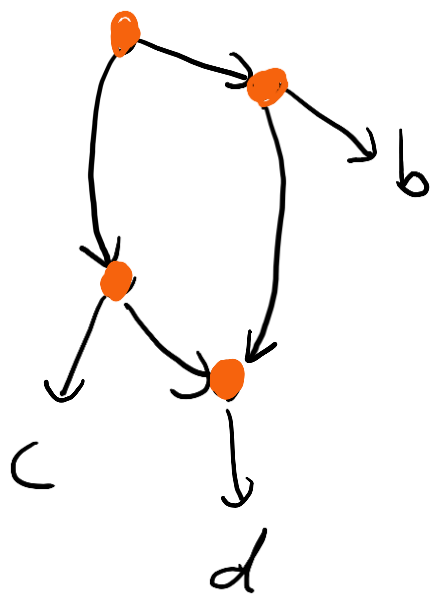
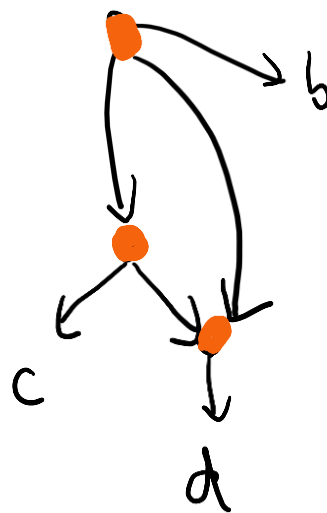
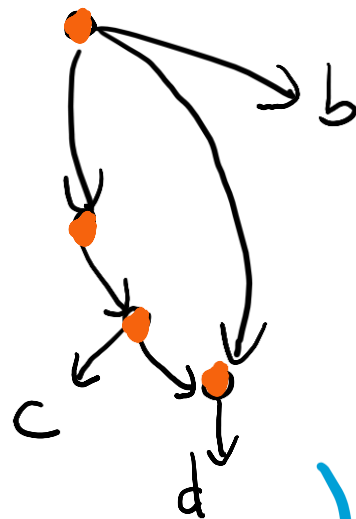
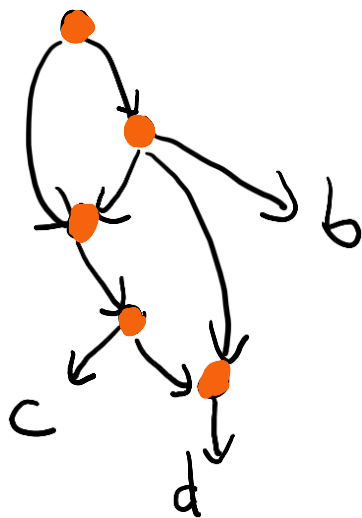
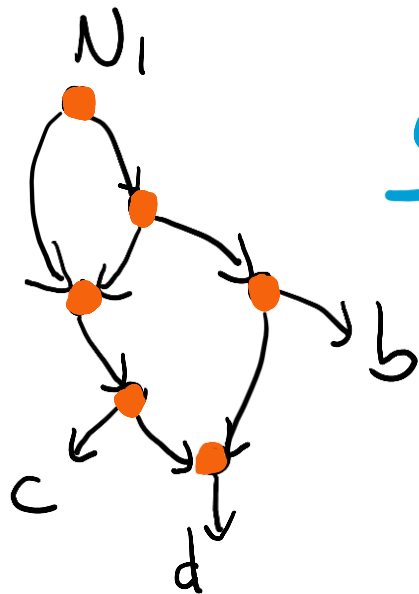




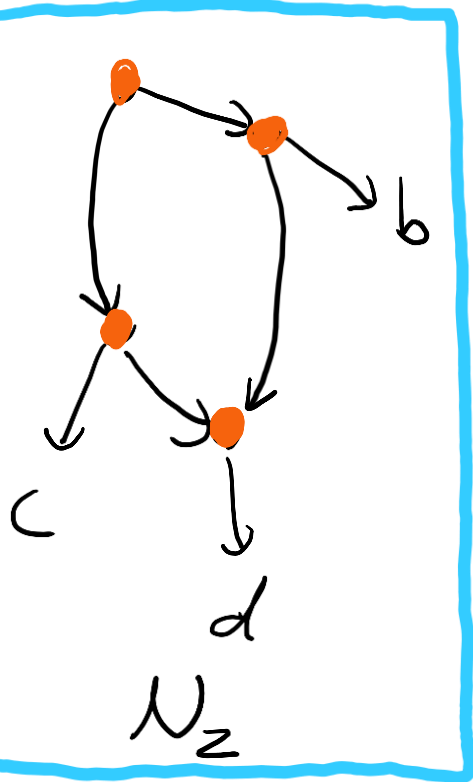
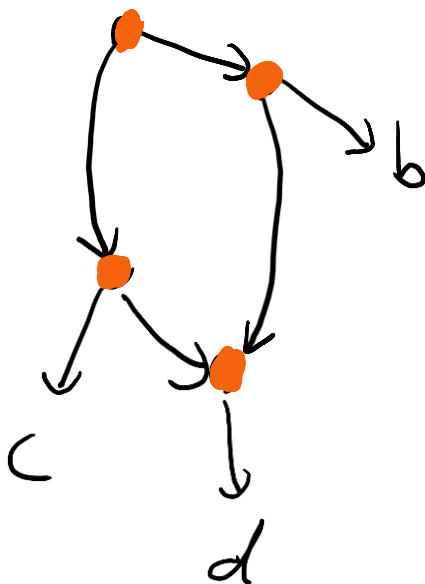
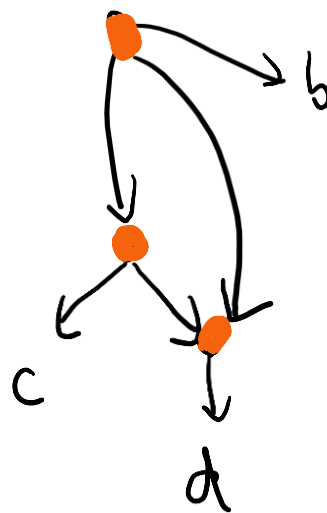
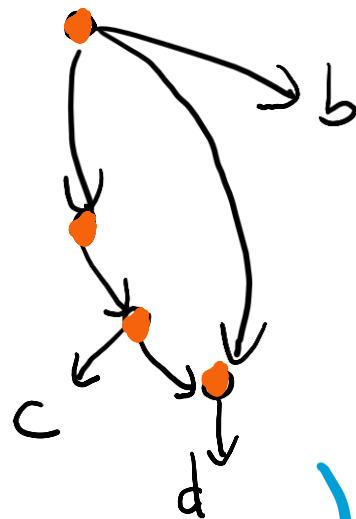
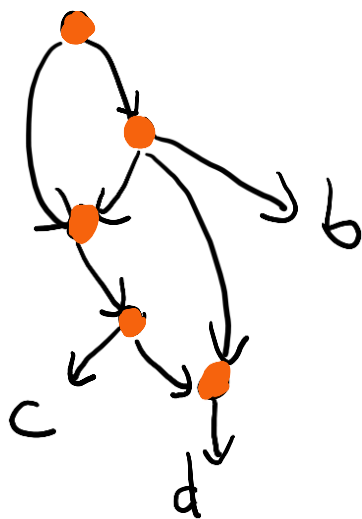
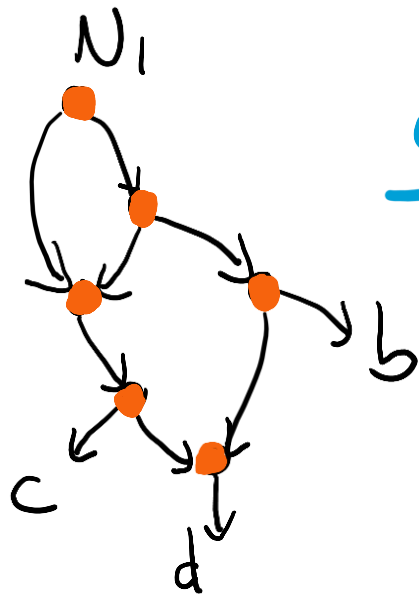
N_2







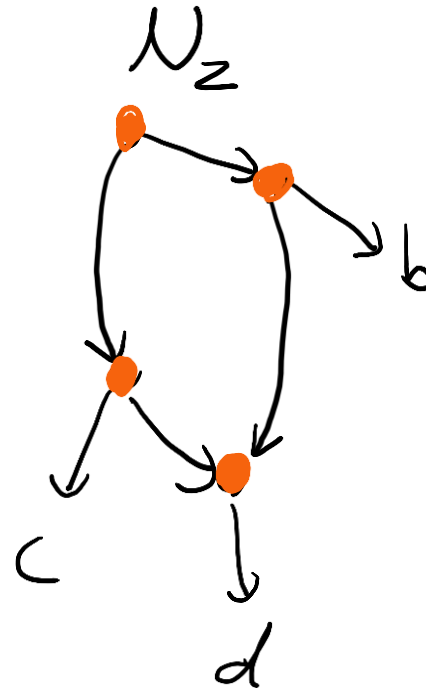
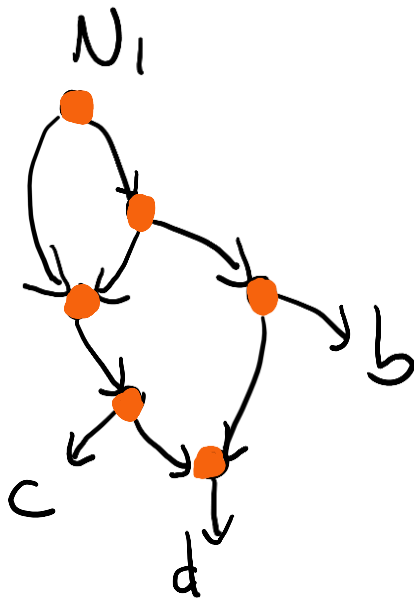
N_2

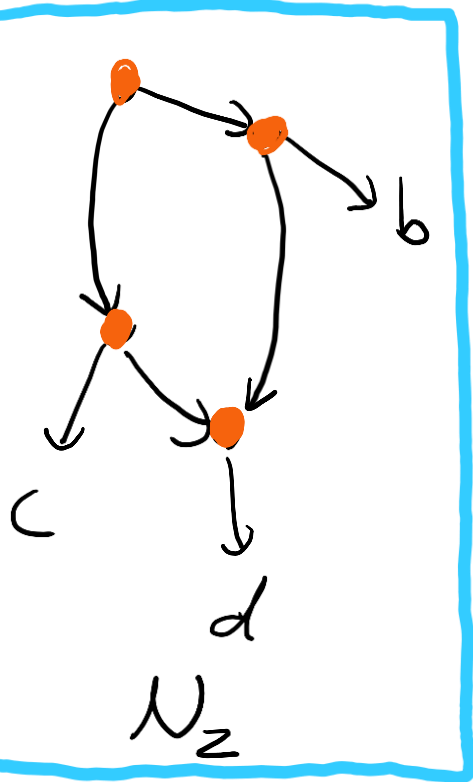
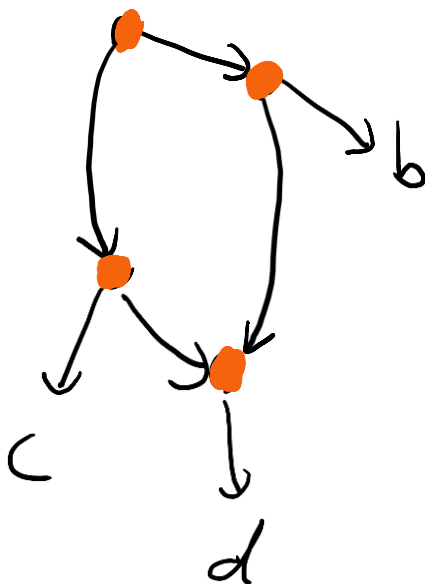
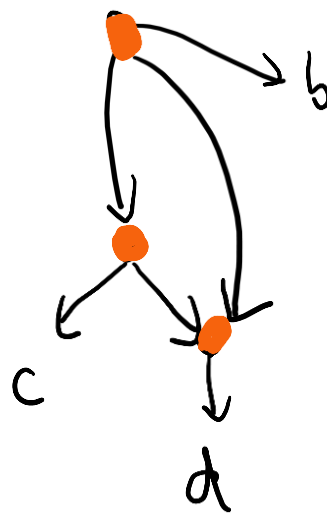
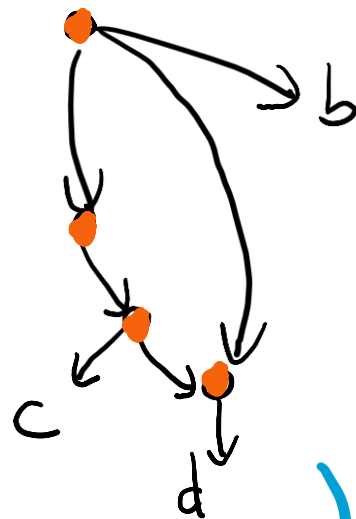
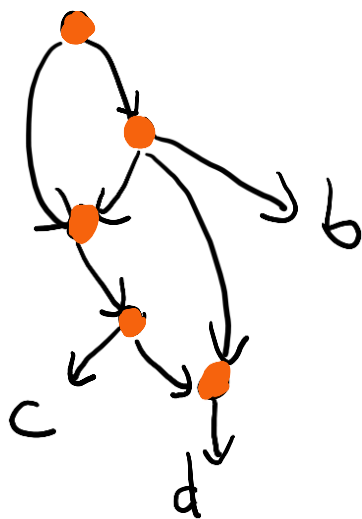
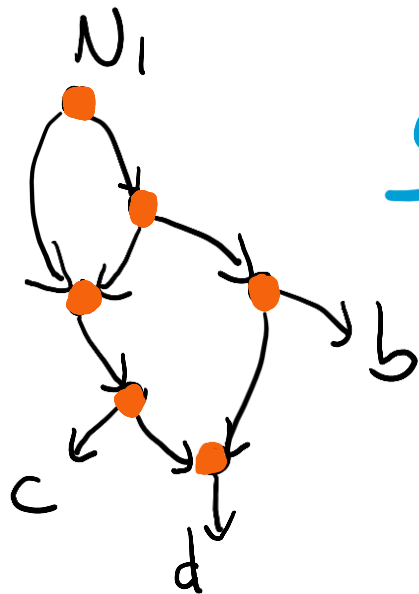


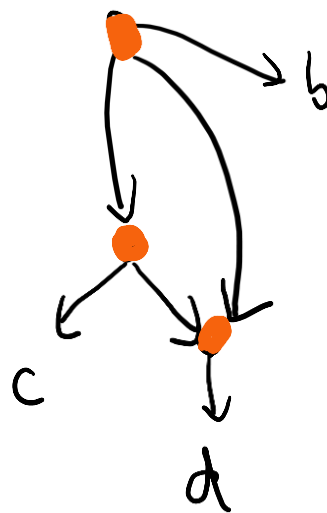
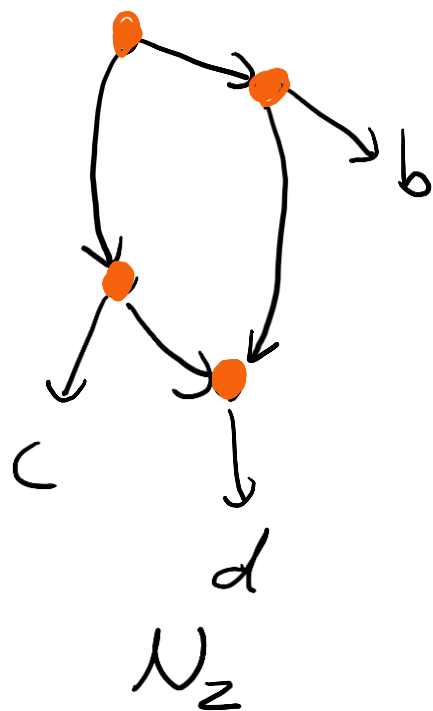
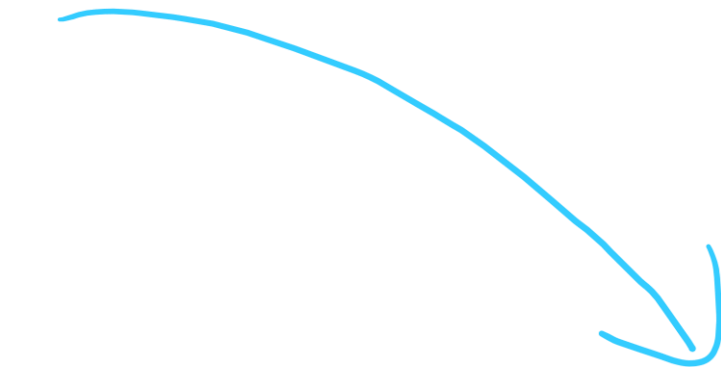
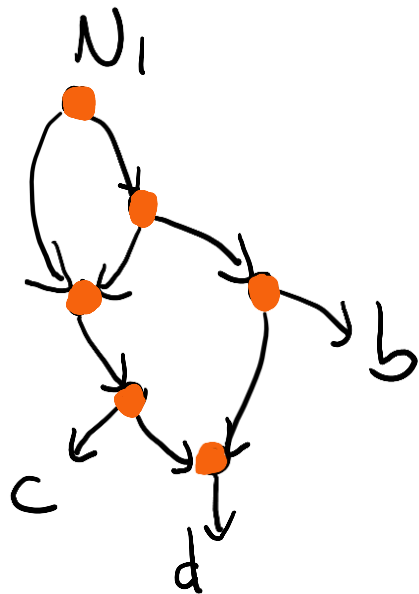
Maximum Common Contraction d_{MCC}

$$d_{MCC}(N_1, N_2) = |N_1| + |N_2| - 2|M|$$

where M is a contraction of both N_1 and N_2 of maximum size.





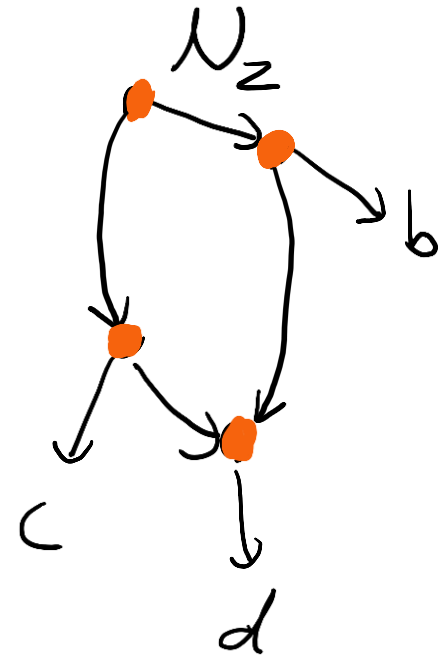
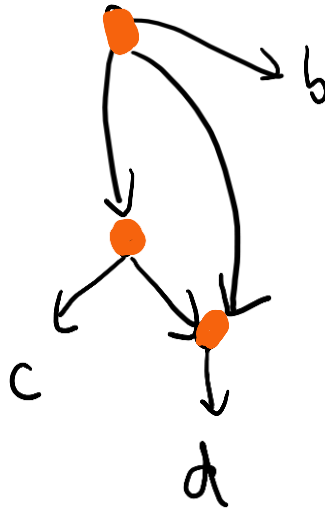
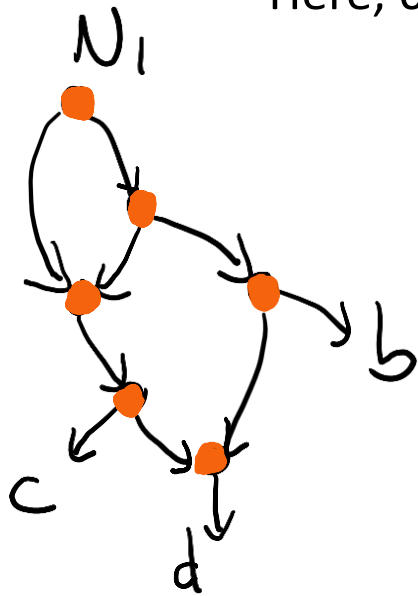


Maximum Common Contraction d_{MCC}

$$d_{MCC}(N_1, N_2) = |N_1| + |N_2| - 2|M|$$

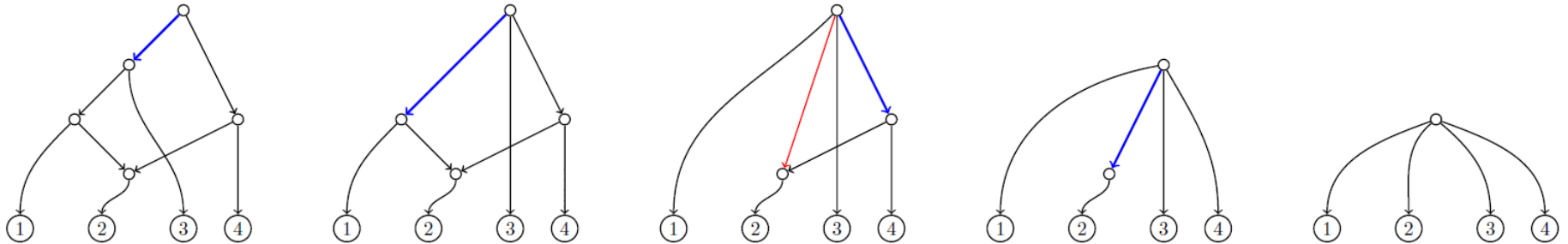
where M is a contraction of both N_1 and N_2 of maximum size.

Here, $6 + 4 - 2 \cdot 3 = 4$ (same as d_{CE})



Contractions + expansions connect the space of **all** networks on the same leafset

Proposition: any network can be contracted into a star (even when considering inadmissible edges). Thus, the star can be expanded into any network.



Contractions connect the space of **all** networks on the same leafset

Proposition: any network can be contracted into a star (even when considering inadmissible edges). Thus, the star can be expanded into any network.

$\Rightarrow$ Any network can be transformed into any network using contractions/expansions.

$$\Rightarrow d_{CE} \leq |N_1| + |N_2| - 2.$$

$$\Rightarrow d_{MCC} \leq |N_1| + |N_2| - 2.$$

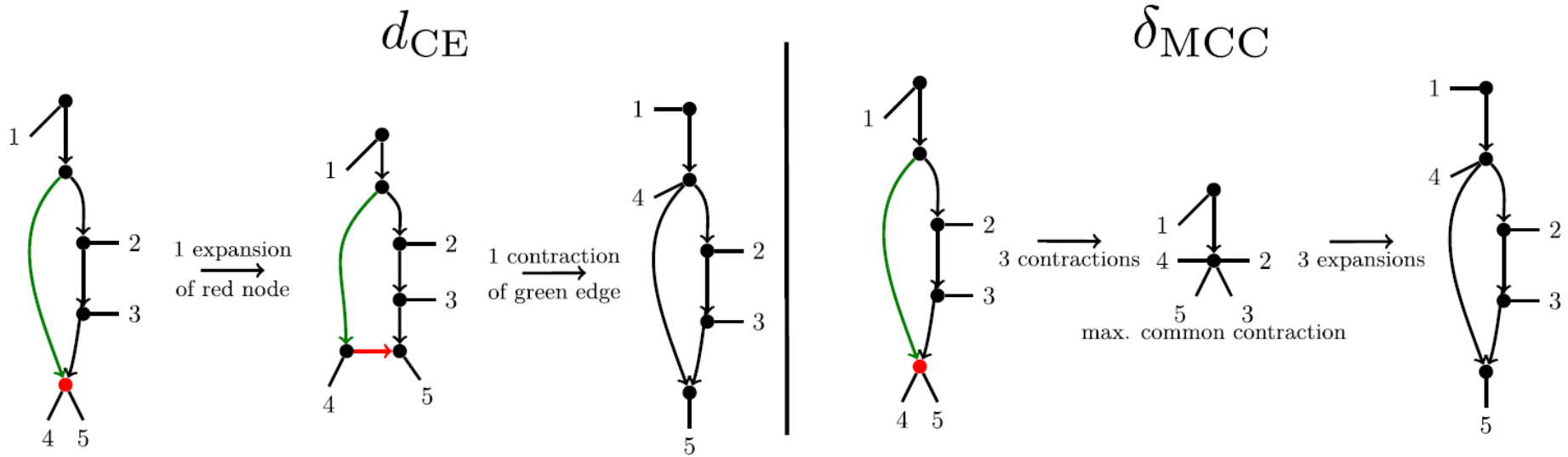
On trees, $d_{CE}(N_1, N_2) = d_{MCC}(N_1, N_2)$

On trees, $d_{CE}(N_1, N_2) = d_{MCC}(N_1, N_2)$

On networks, $d_{CE}(N_1, N_2) \neq d_{MCC}(N_1, N_2)$ is possible

On trees, $d_{CE}(N_1, N_2) = d_{MCC}(N_1, N_2)$

On networks, $d_{CE}(N_1, N_2) \neq d_{MCC}(N_1, N_2)$ is possible



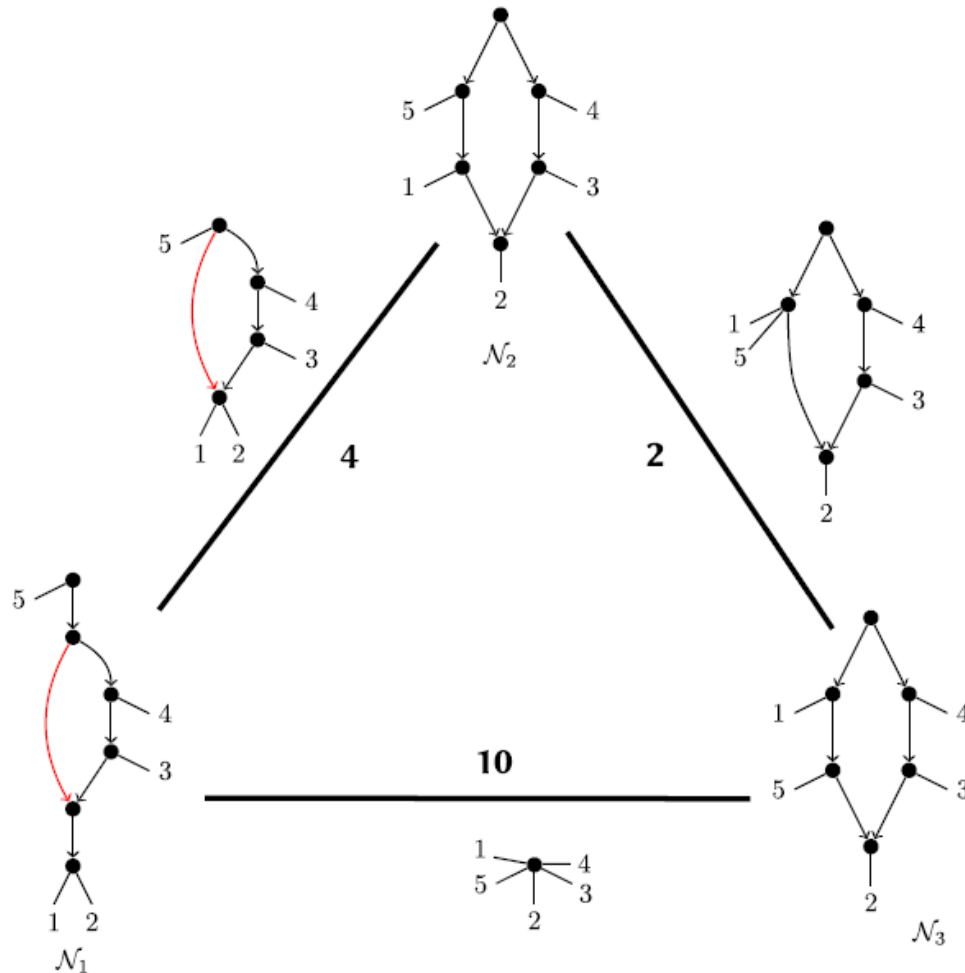
d_{CE} is a metric

- Identity: $d(N_1, N_1) = 0$ trivially
- Positivity: $N_1 \neq N_2$ implies $d(N_1, N_2) > 0$: by definition
- Symmetry: $d(N_1, N_2) = d(N_2, N_1)$ since contractions/expansions are the reverse of each other
- Triangle inequality: $d(N_1, N_3) \leq d(N_1, N_2) + d(N_2, N_3)$

Just transform N_1 into N_2 , then N_2 into N_3 .

d_{MCC} is a semi-metric $|N_1| + |N_2| - 2|M|$

- It does not satisfy the triangle inequality.



Diameter

- F_n = set of all networks with n internal vertices
- Diameter is defined w.r.t. size of your N_1 and N_2 of interest:

$$\text{diam}(n, m) = \max_{N_1 \in F_n, N_2 \in F_m} \delta_{MCC}(N_1, N_2)$$

Proposition: $\text{diam}(n, m) = n + m - 2$

$\Rightarrow$ for any n, m , there are instances in which we must contract to a star tree.

Algorithmic aspects

The Maximum Common Contraction Problem

- **Input:** networks N_1, N_2 on the same leafset, integer k .
- **Question:** is there a common contraction M with at least k vertices.

The Maximum Common Contraction Problem

- **Input:** networks N_1, N_2 on the same leafset, integer k .
- **Question:** is there a common contraction M with at least k vertices.

Results

- NP-hard even on networks with maximum degree 10. ☹️
- NP-hard even if N_1, N_2 have 5 leaves and N_2 is a tree. ☹️
- Positive: solvable in time $O(n^5)$ on weakly galled trees. 😊

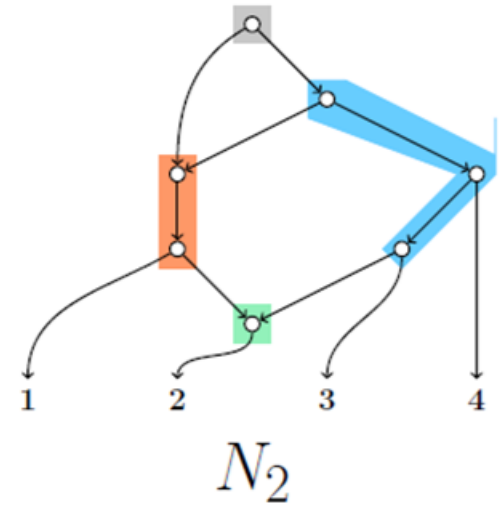
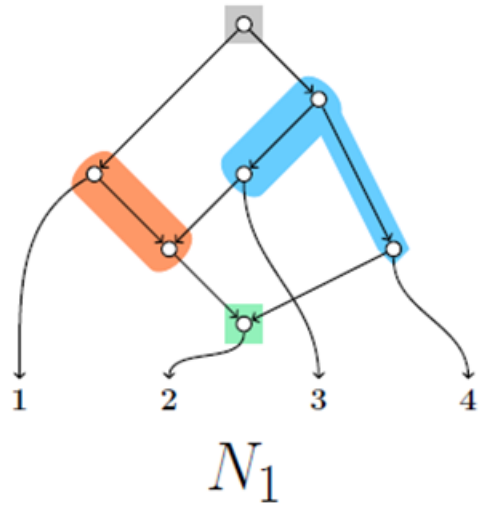
The Maximum Common Contraction Problem

- **Input:** networks N_1, N_2 on the same leafset, integer k .
- **Question:** is there a common contraction M with at least k vertices.

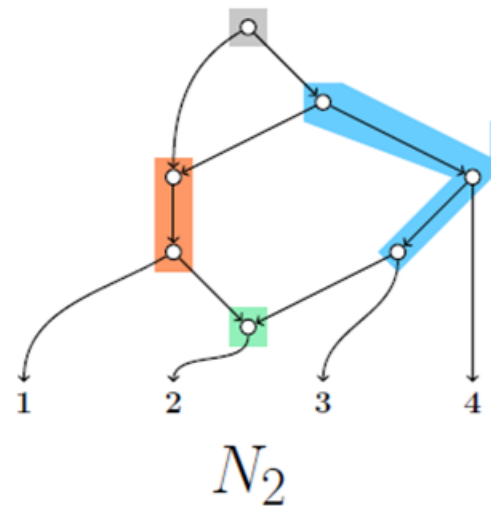
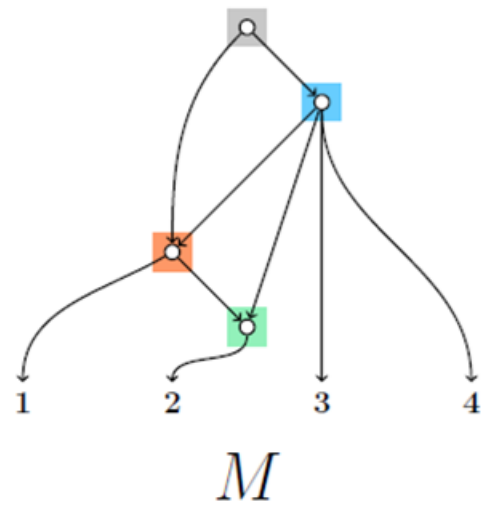
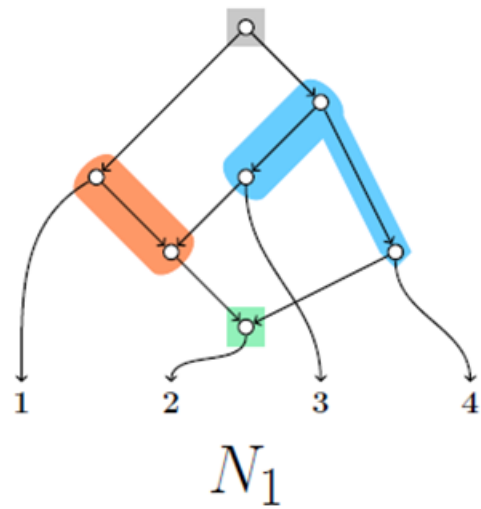
Results

- NP-hard even on networks with maximum degree 10. ☹️
- NP-hard even if N_1, N_2 have 5 leaves and N_2 is a tree. ☹️
 - para-NP-hard in max degree, in number of leaves, in $|N_2|$.
 - no subexponential time algorithm if Exponential Time Hypothesis (ETH) is true.
- Positive: solvable in time $O(n^5)$ on weakly galled trees. 😊

Some intuition: **witness-structure**



Some intuition: **witness-structure**



Set Splitting reduces to MCC

- Input: subsets $A = \{A_1, A_2, \dots, A_m\}$ of $X = \{1, 2, 3, \dots, n\}$.
- Task: color X in two colors so that no A_i is monochromatic.

A : 1,2,3 2,3,4 1,3 1,4

Set Splitting reduces to MCC

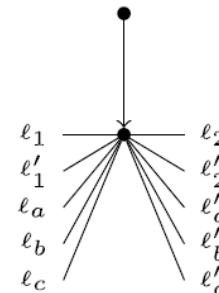
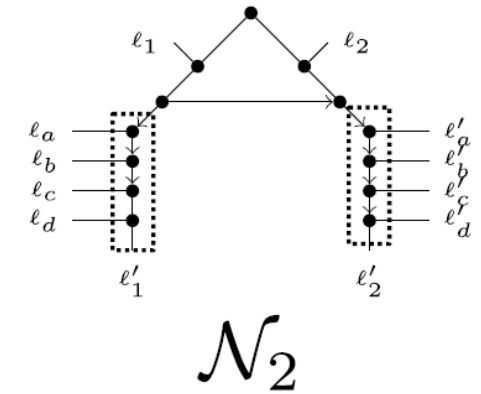
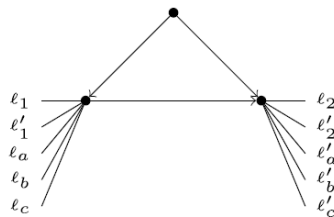
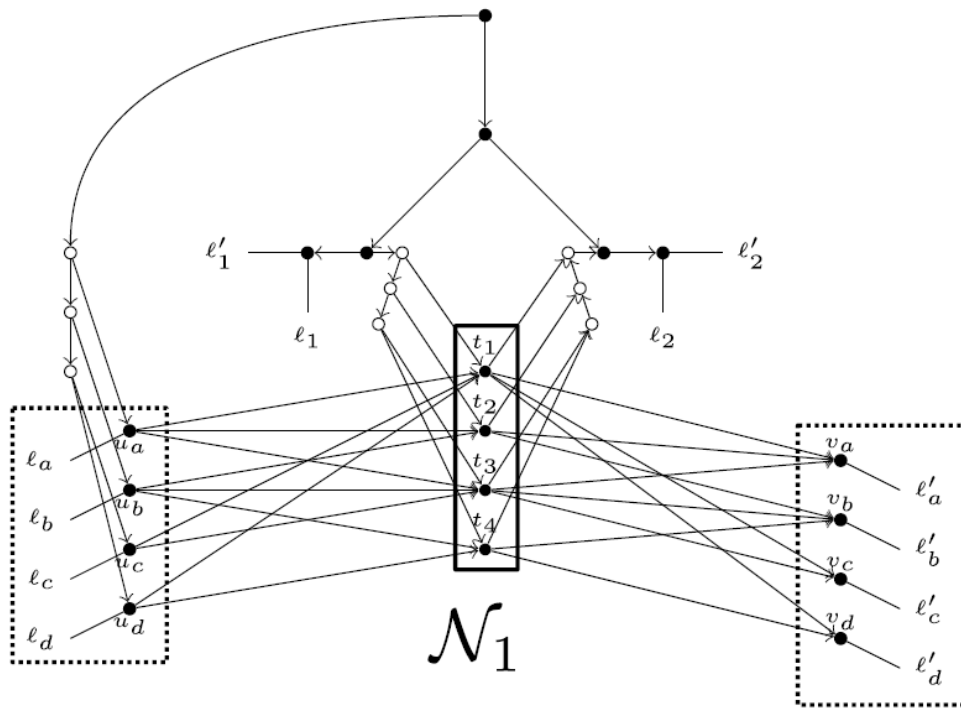
- Input: subsets $A = \{A_1, A_2, \dots, A_m\}$ of $X = \{1, 2, 3, \dots, n\}$.
- Task: color X in two colors so that no A_i is monochromatic.

A :  $\{1, 2, 3, 4\}$

Set Splitting reduces to MCC

- Input: subsets $A = \{A_1, A_2, \dots, A_m\}$ of $X = \{1, 2, 3, \dots, n\}$.
- Task: color X in two colors so that no A_i is monochromatic.

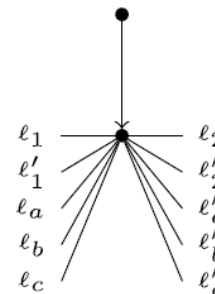
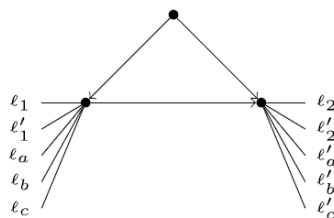
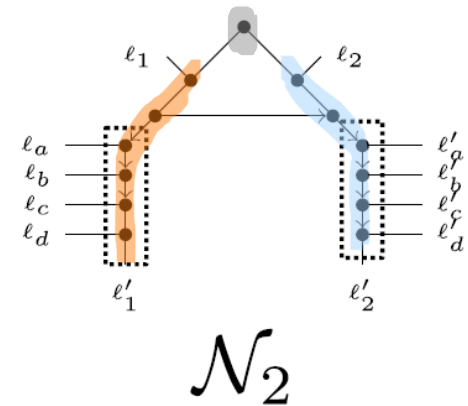
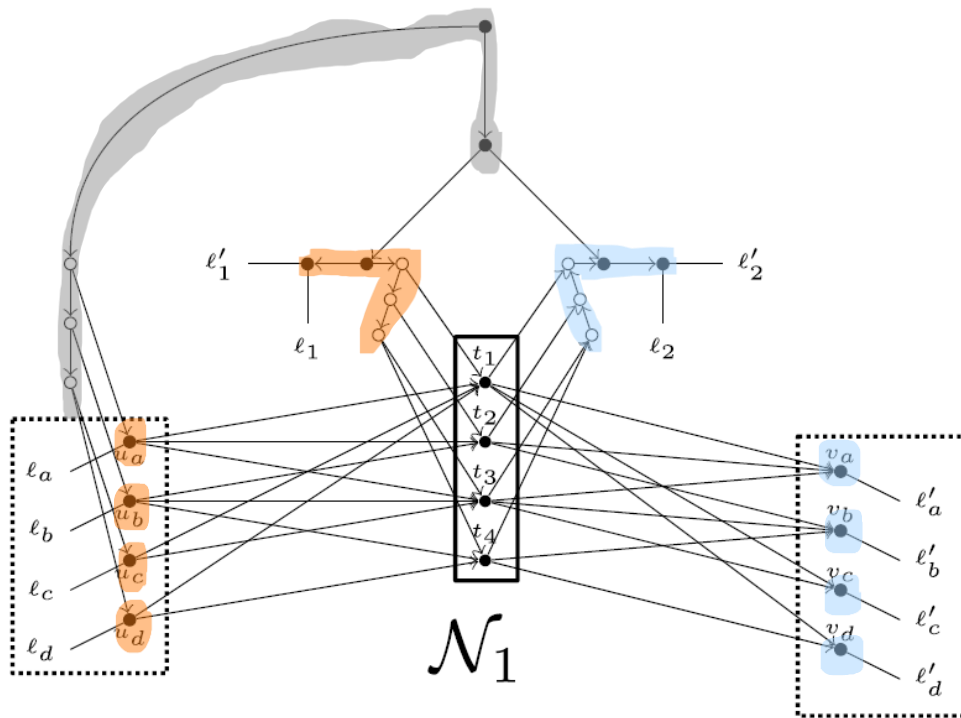
A : 1,2,3 2,3,4 1,3 1,4



Set Splitting reduces to MCC

- Input: subsets $A = \{A_1, A_2, \dots, A_m\}$ of $X = \{1, 2, 3, \dots, n\}$.
- Task: color X in two colors so that no A_i is monochromatic.

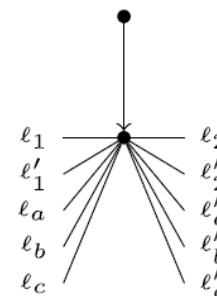
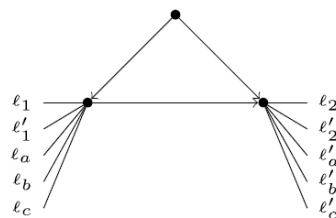
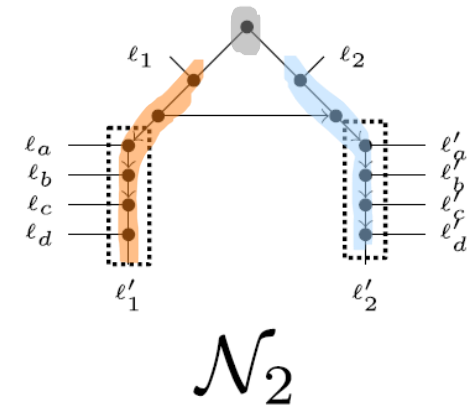
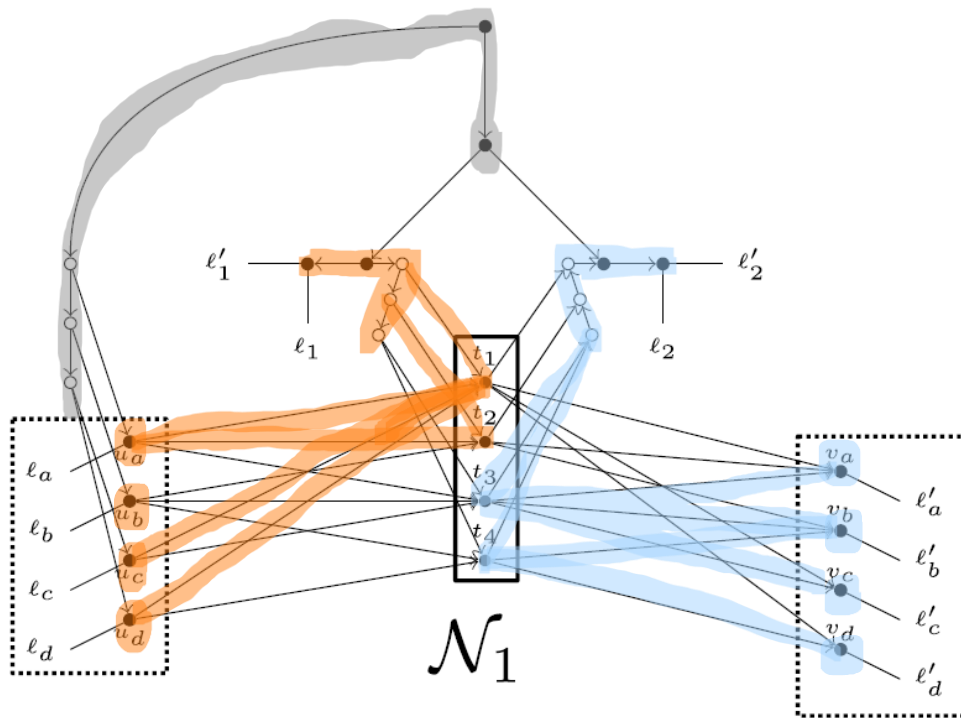
A : 1,2,3 2,3,4 1,3 1,4



Set Splitting reduces to MCC

- Input: subsets $A = \{A_1, A_2, \dots, A_m\}$ of $X = \{1, 2, 3, \dots, n\}$.
- Task: color X in two colors so that no A_i is monochromatic.

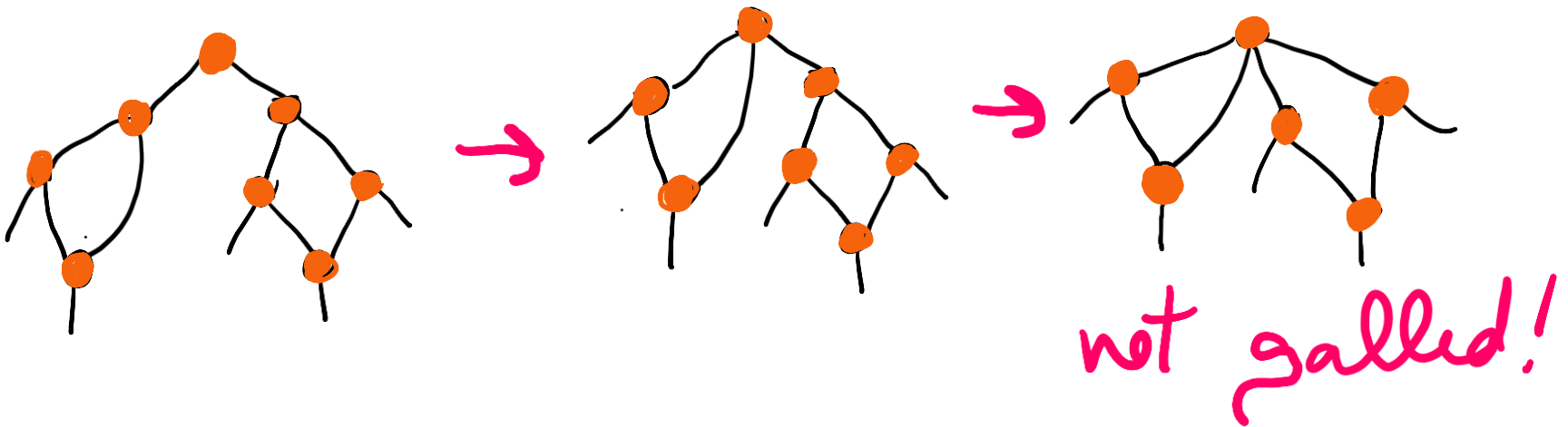
A : 1, 2, 3 2, 3, 4 1, 3 1, 4



NP-hard on general networks => focus on network classes

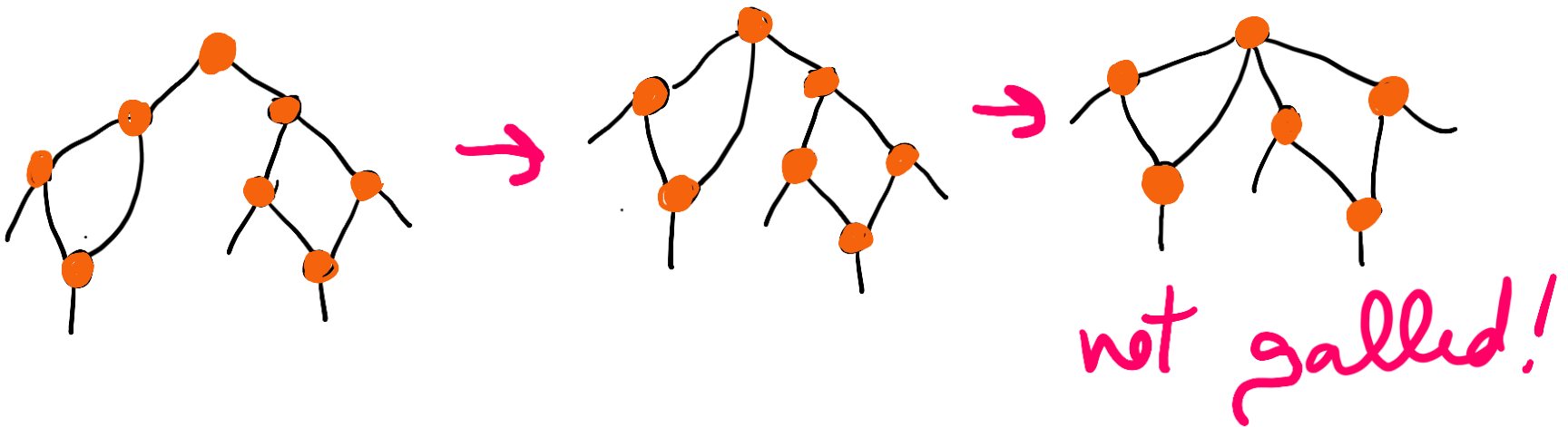
NP-hard on general networks => focus on network classes

- Start with the simplest class beyond trees: galled trees (cycles are vertex-disjoint)
- Strategy:
 1. find a contraction that MUST be applied on either N_1 or N_2
 2. if such a contraction exists, apply it and repeat
 3. use structural properties of networks with no forced contractions



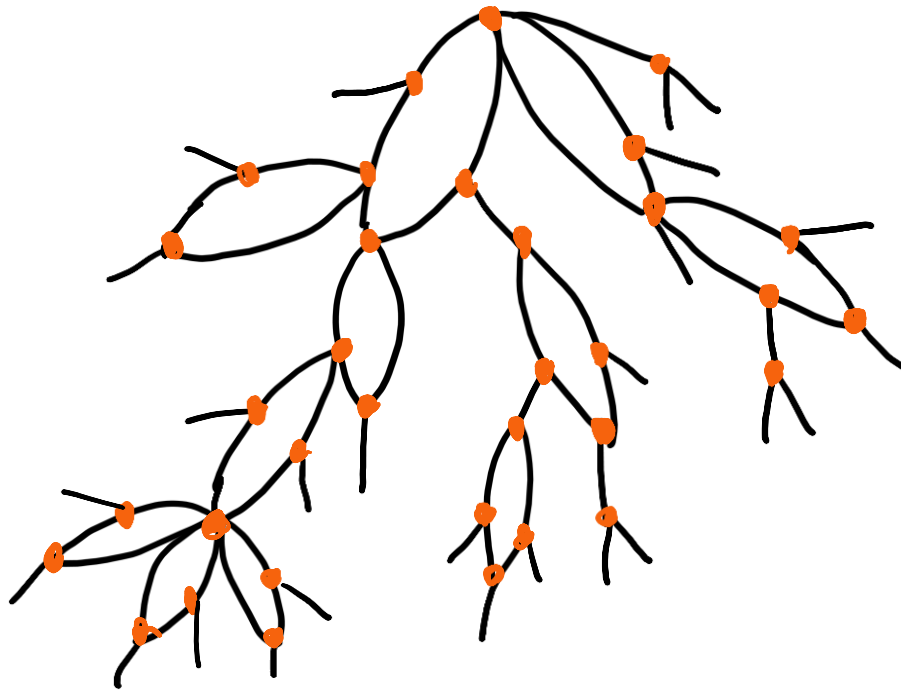
NP-hard on general networks => focus on network classes

- Start with the simplest class beyond trees: galled trees (cycles are vertex-disjoint)
- Strategy:
 1. find a contraction that **MUST** be applied on either N_1 or N_2
 2. if such a contraction exists, apply it and repeat
 3. use structural properties of networks with no forced contractions
- Problem: after applying one contraction, we can be out of the class of galled trees.
- Solution: find the simplest class that is **closed under contractions**.



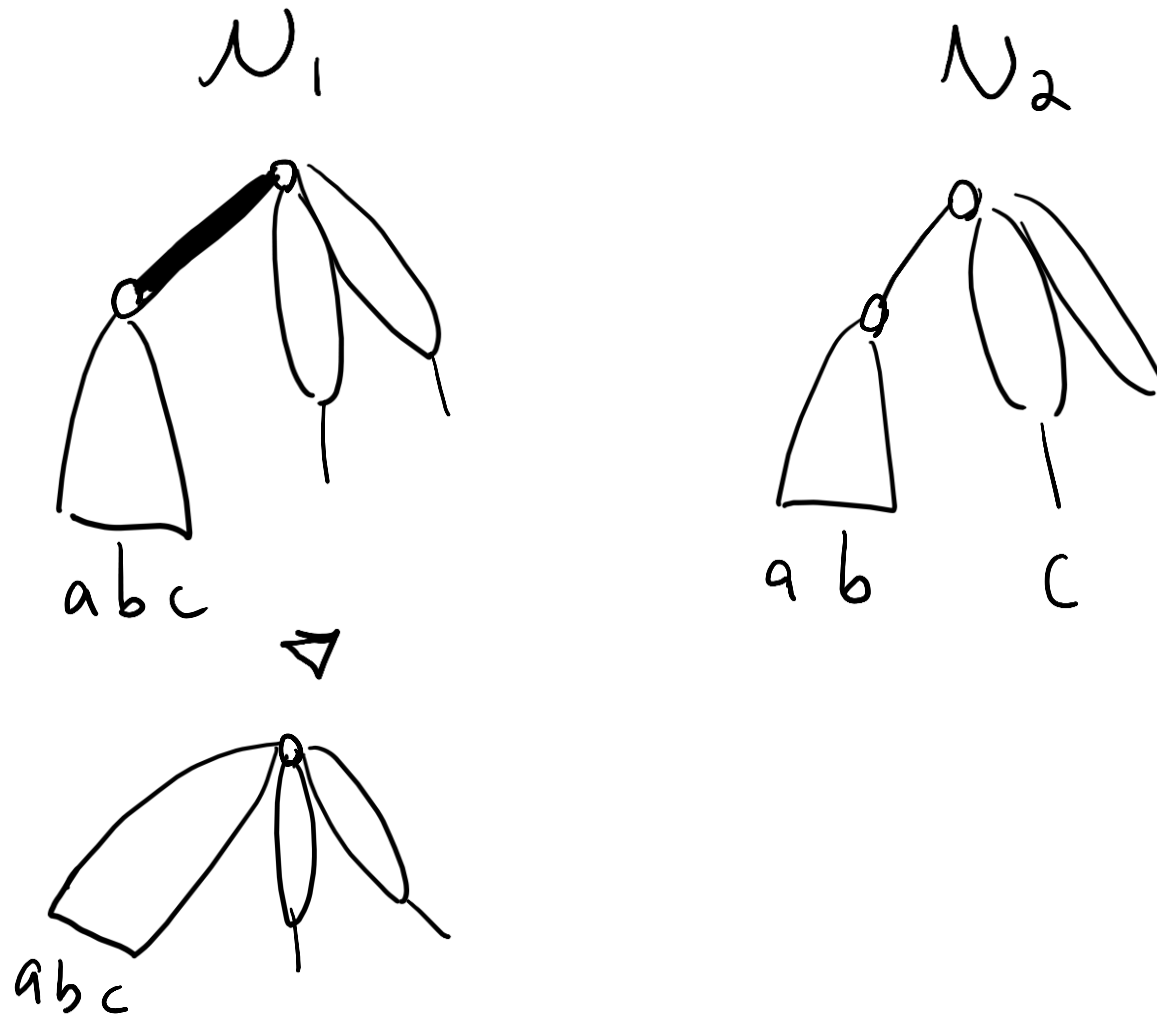
Weakly galled trees

- Networks in which each arc is contained in at most one cycle
 - Cycles do not have edges in common (but possibly vertices)
 - Essentially level-1, but reticulations have in-degree 2, not more.
 - (from the classification of [Rosselló, Valiente, 2009])

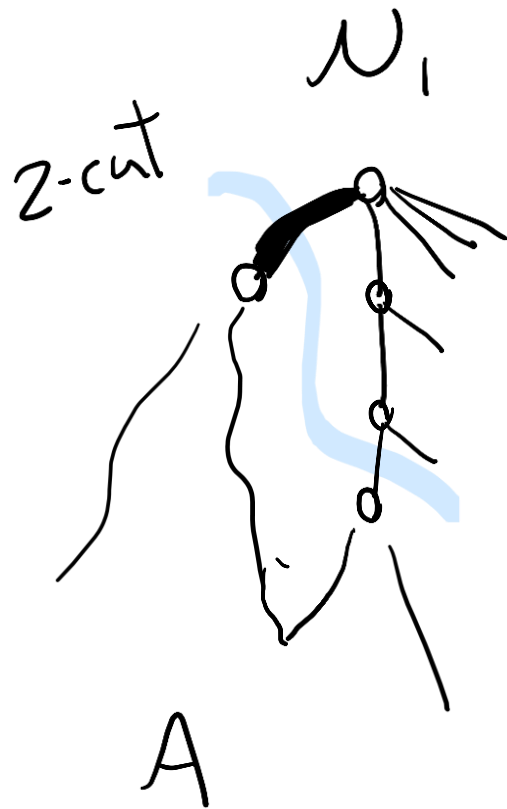


(aka cactus DAGs)

- Look at the arcs incident to the root which MUST be contracted.
 - Tree arcs whose cluster is not in the other network
 - Cycle arcs whose 2-cuts are not in the other network



- Look at the arcs incident to the root which MUST be contracted.
 - Tree arcs whose cluster is not in the other network
 - Cycle arcs whose 2-cuts are not in the other network

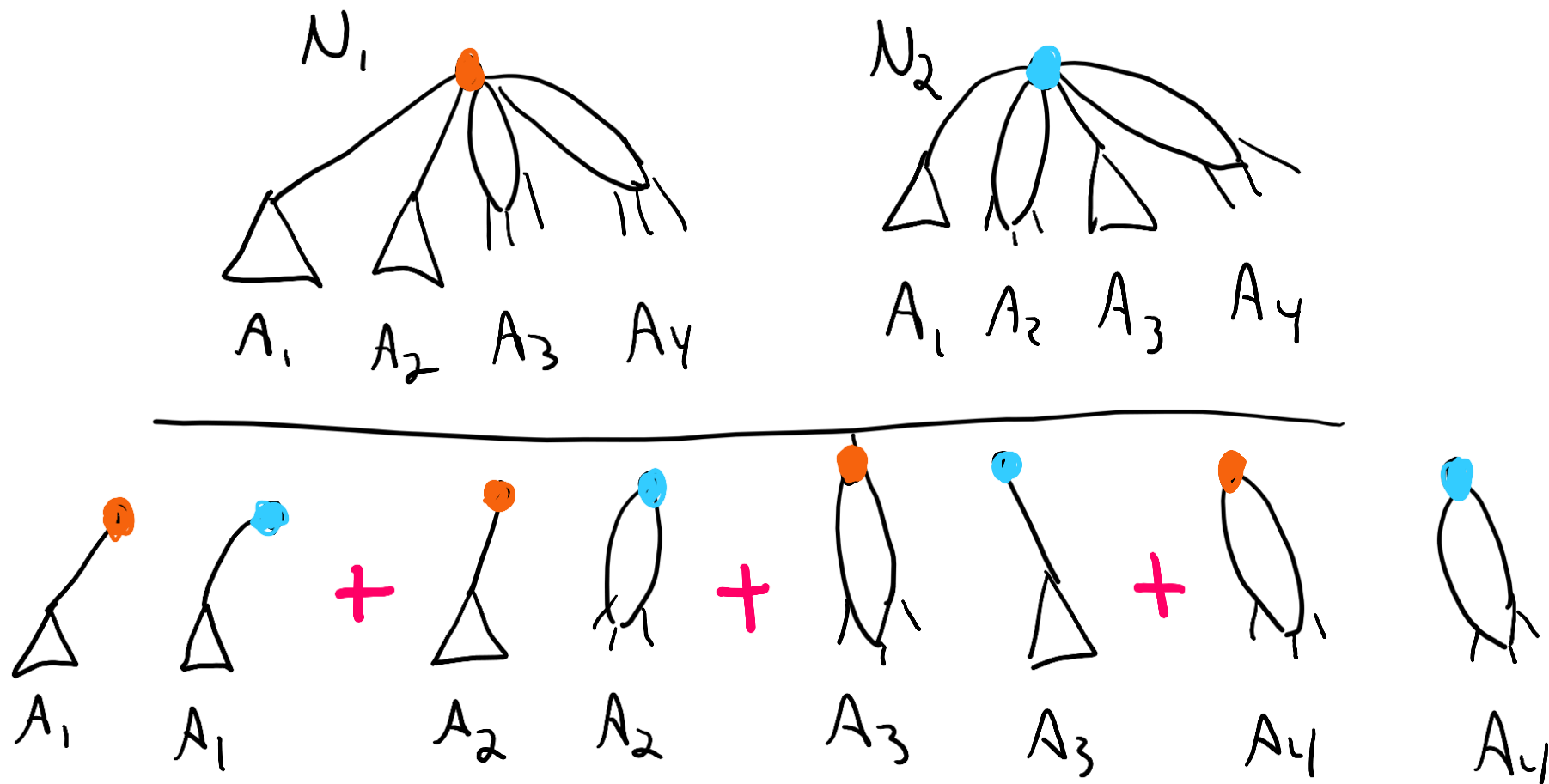


N_2

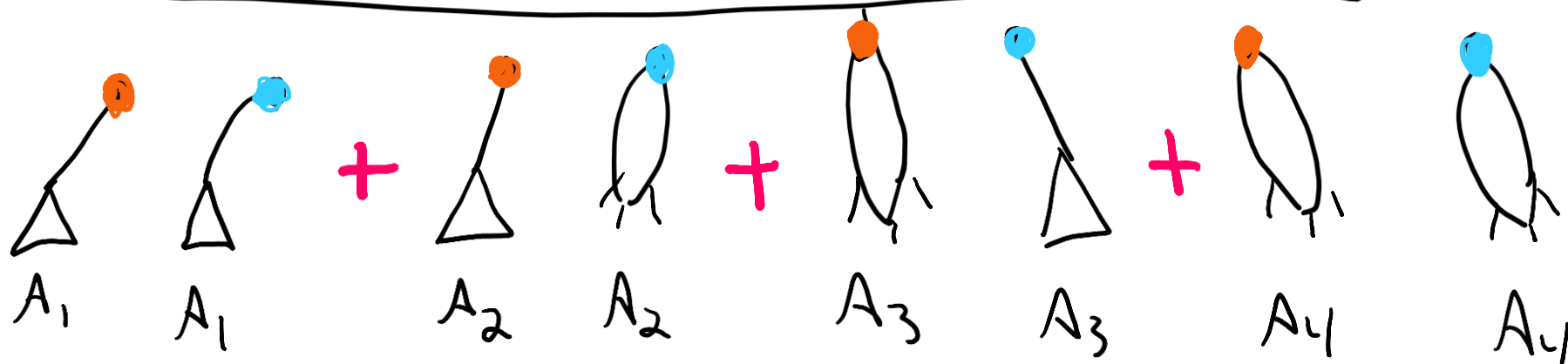
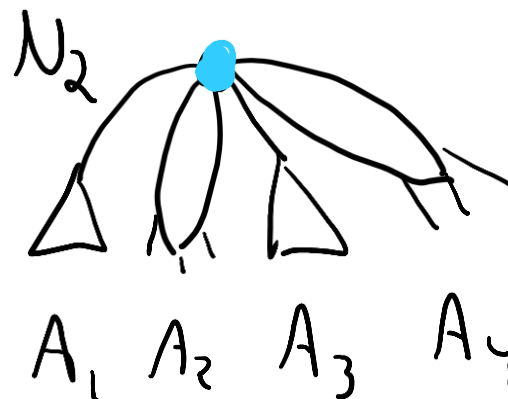
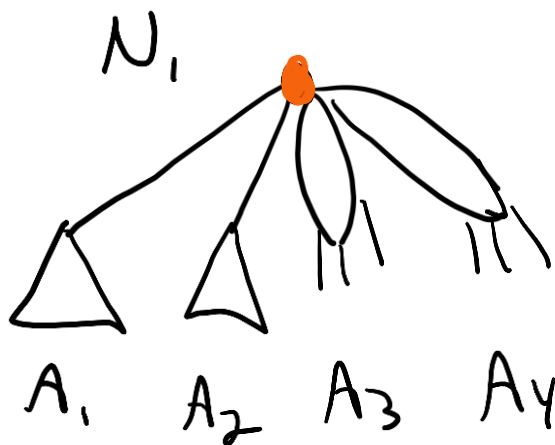
no 2-cut
splits into $A | \bar{A}$

- A weakly galled tree is *simple* if its root has outdegree 1 or 2.
- **Proposition:** if N_1, N_2 have no forced contractions at the root r , then we can decompose N_1, N_2 into pairs of simple networks with the same leafsets.
- Problem is just to compare simple networks => dynamic programming

- A weakly galled tree is *simple* if its root has outdegree 1 or 2.
- **Proposition:** if N_1, N_2 have no forced contractions at the root r , then we can decompose N_1, N_2 into pairs of simple networks with the same leafsets.
- Problem is just to compare simple networks => dynamic programming



- A weakly galled tree is *simple* if its root has outdegree 1 or 2.
- **Proposition:** if N_1, N_2 have no forced contractions at the root r , then we can decompose N_1, N_2 into pairs of simple networks with the same leafsets.
- Problem is just to compare simple networks => dynamic programming



- **Open problem:** is d_{MCC} fixed-parameter tractable in the **level**? Or in the hybridization number? Or any useful parameter?
- **Vague open problem 1:** design the perfect metric on phylogenetic networks.
 - Or just an operational metric that is easy to compute.
- **Vague open problem 2:** determine whether contractions/expansions can be useful to explore the space of networks.