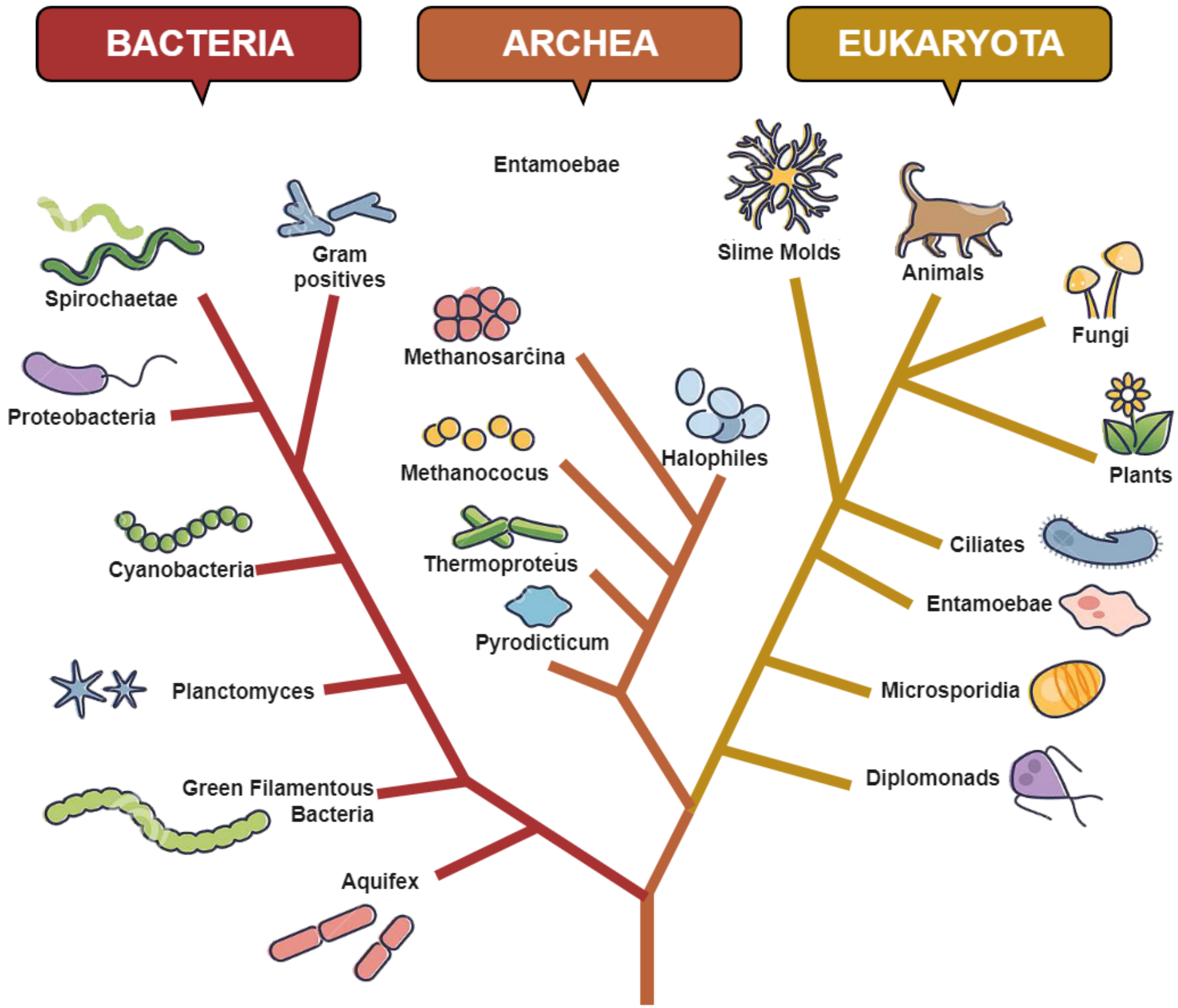
The $\ominus$-metric to compare phylogenetic networks

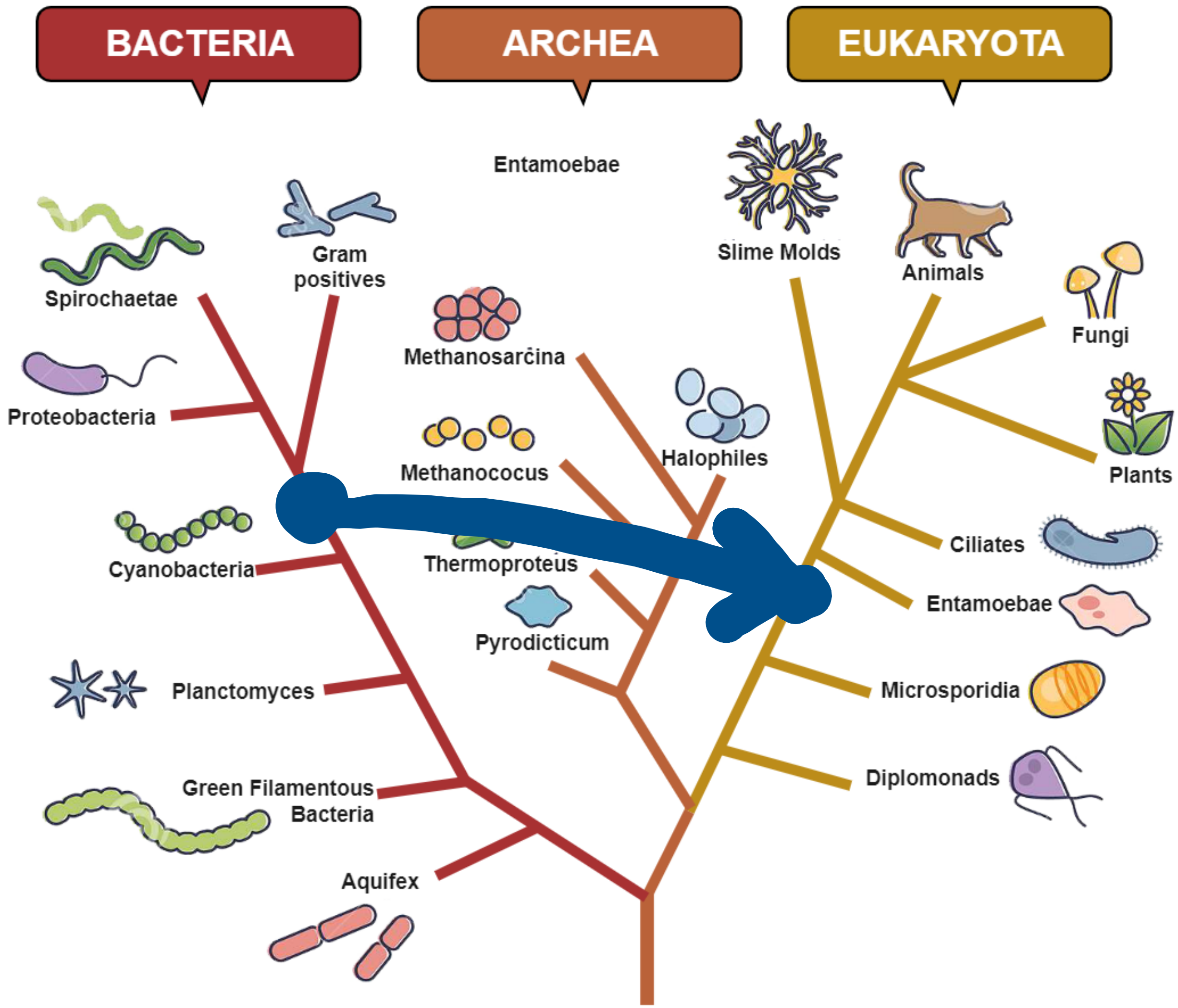
Marc Hellmuth Leipzig University, Germany

Manuel Lafond Université de Sherbrooke, Canada

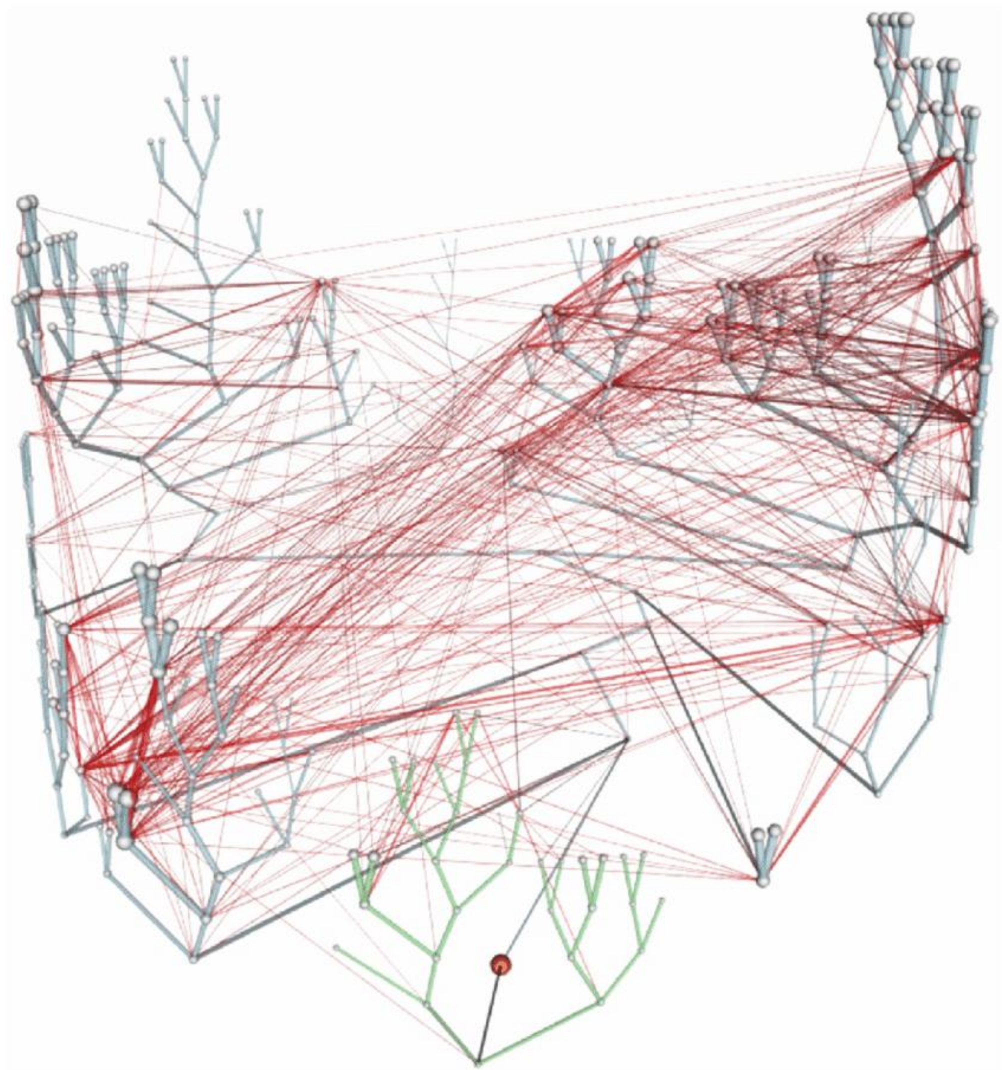
Guillaume Scholz University of Greifswald, Germany



BACTERIA
ARCHEA
EUKARYOTA
Entamoebae
Spirochaetae
Gram positives
Proteobacteria
Cyanobacteria
Planctomyces
Green Filamentous Bacteria
Aquifex
Methanosarĉina
Halophiles
Methanococus
Thermoproteus
Pyrodicticum
Slime Molds
Animals
Fungi
Plants
Ciliates
Entamoebae
Microsporidia
Diplomonads



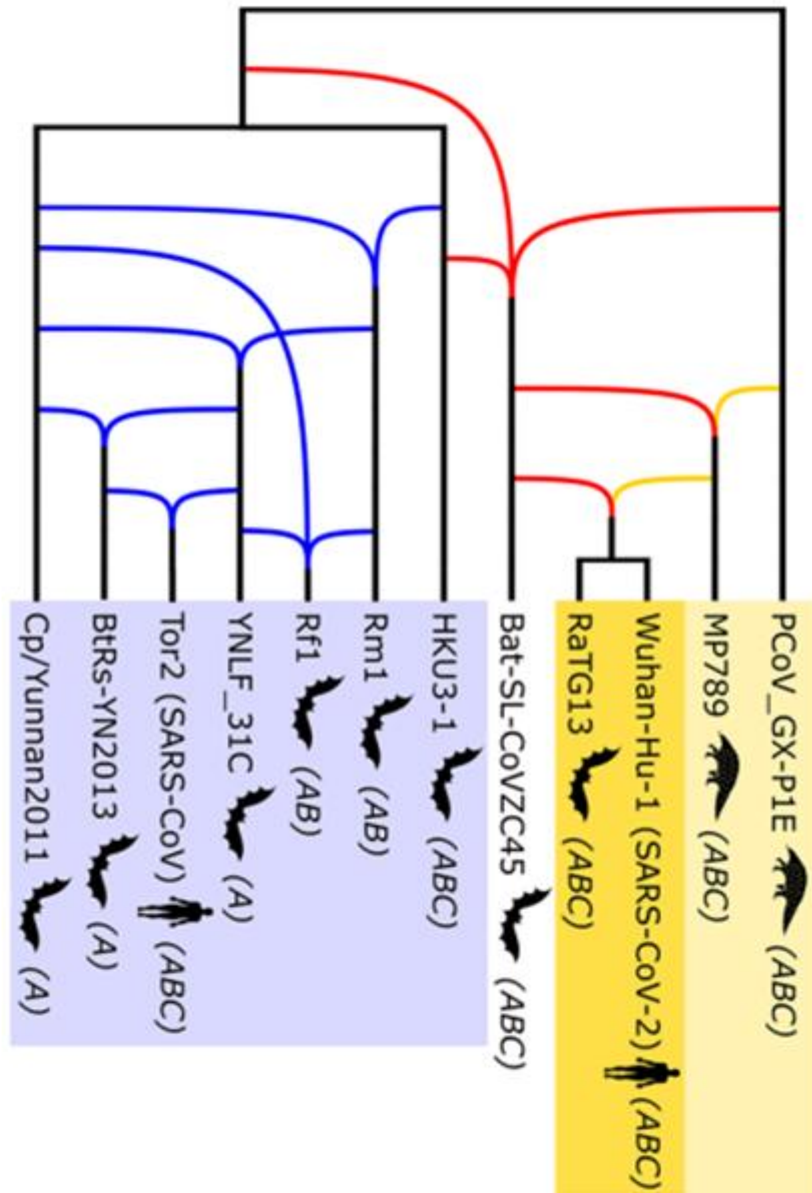
BACTERIA
ARCHEA
EUKARYOTA
Entamoebae
Spirochaetae
Gram
positives
Proteobacteria
Cyanobacteria
Planctomyces
Green Filamentous
Bacteria
Aquifex
Methanosarĉina
Methanococus
Halophiles
Thermoproteus
Pyrodicticum
Slime Molds
Animals
Fungi
Plants
Ciliates
Entamoebae
Microsporidia
Diplomonads



RESEARCH Open Access

Applicability of several rooted phylogenetic network algorithms for representing the evolutionary history of SARS-CoV-2

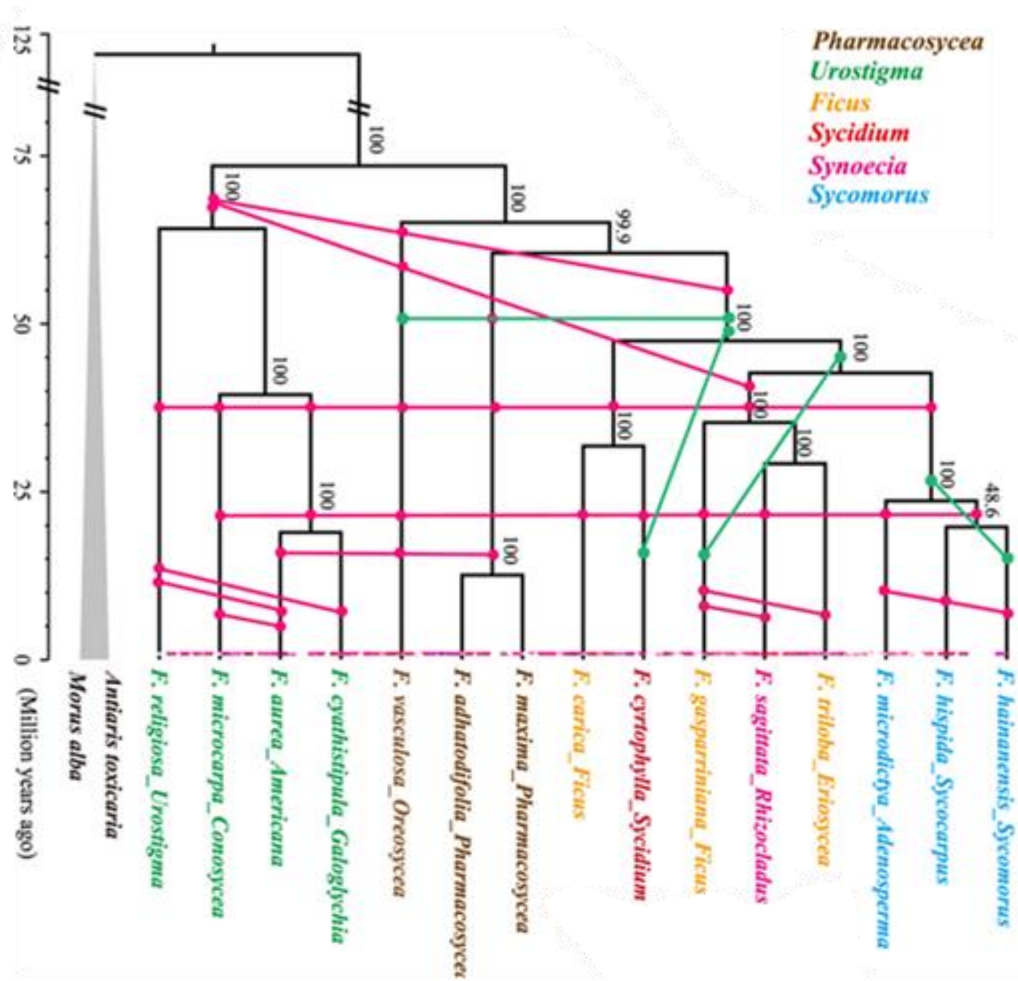
Rosanne Wallin[1], Leo van Iersel[2], Steven Kelk[3] and Leen Stougie[1,4*]



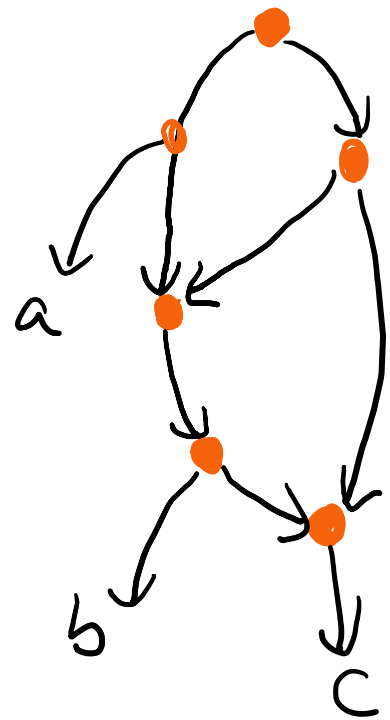
Genomic evidence of prevalent hybridization throughout the evolutionary history of the fig-wasp pollination mutualism

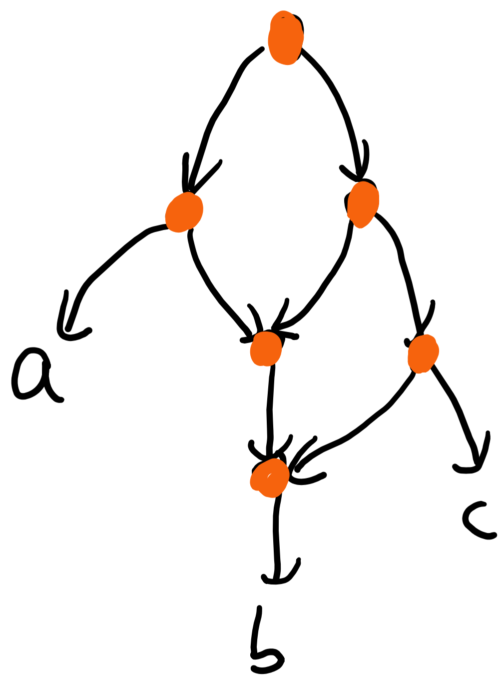
Gang Wang, Xingtan Zhang, Edward Allen Herre, Doyle McKey, Carlos A. Machado, Wen-Bin Yu, Charles H. Cannon, Michael L. Arnold, Rodrigo A. S. Pereira, Ray Ming, Yi-Fei Liu, Yibin Wang, Dongna Ma & Jin Chen

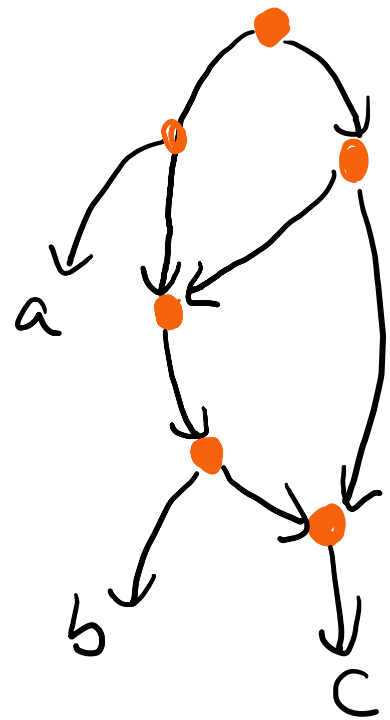
Nature Communications **12**, Article number: 718 (2021) | Cite this article




THE MAGIC MACHINE

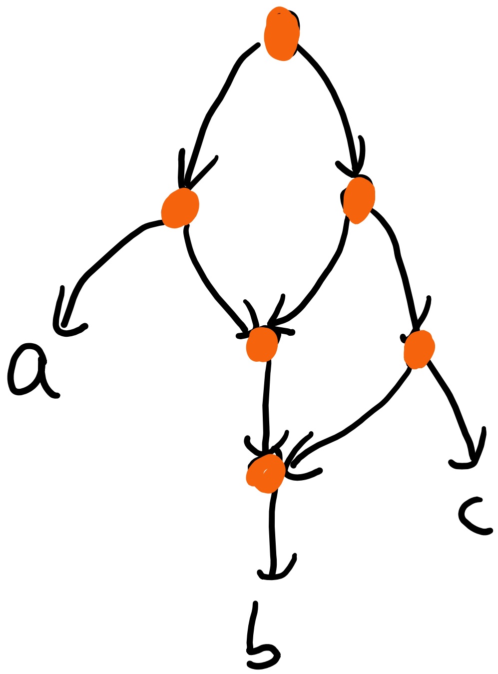

a
b
c

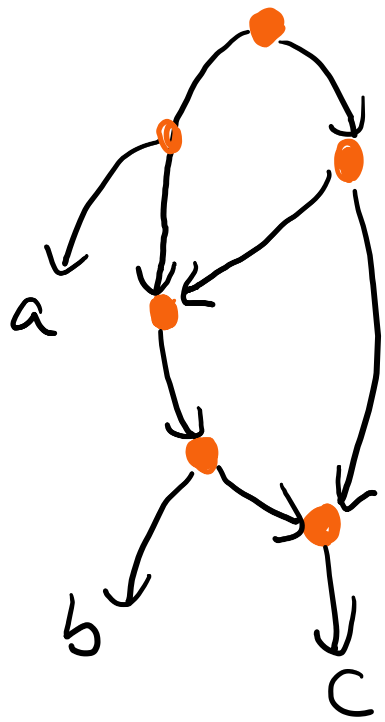

a
b
c


a
b
c

Metrics and dissimilarity measures on networks:

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





Why metrics and dissimilarity measures

- Compare inferred networks against gold standard
- Extract common structure.
- Get a consensus.
- 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

ARTICLE | ANTHROPOLOGY

...etwork analysis of SARS-CoV-2

LETTER | ANTHROPOLOGY

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

Santiago J. Sánchez-Pacheco, Sungsik Kong, Paola Pulido-Santacruz, +1, and Laura Kubatko Authors Info &
Affiliations

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

ARTICLE | ANTHROPOLOGY

...twork analysis of SARS-CoV-2

LETTER | ANTHROPOLOGY

Median-joining netw...

LETTER | ANTHROPOLOGY

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

twork analysis of SARS-CoV-2

LETTER | ANTHROPOLOGY

Median-joining netw

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

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	is ≠0 if N≠N'	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	$O(2.4^d\ n)$	NO (# retics)	YES
Cherry-picking	YES	$O(3^r\ n^c)$	NO (orchards)	YES
Contractions	YES	NP-hard	YES, but no Δ-ineq.	YES
d_{LGT}	YES	NP-hard, poly variant	NO (LGT networks)	YES

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

Operational	is ≠0 if N≠N'	Compute time	Compare any networks?	Common structure?
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
Contractions	YES	NP-hard	YES, but no Δ-ineq.	YES
d_{LGT}	YES	NP-hard, poly variant	NO (LGT networks)	YES

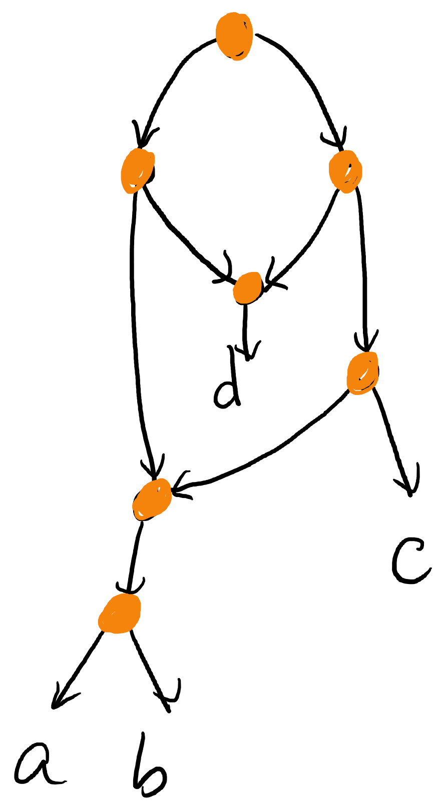
* yes on orchards ** yes on time-consistent tree-child   *** add triangle creation move

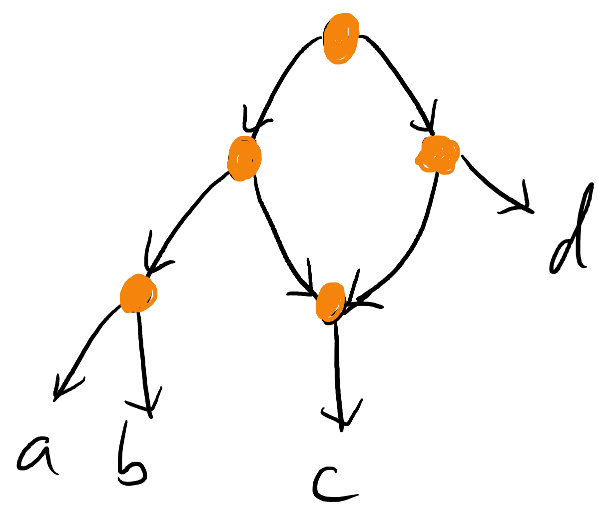
Operational	is ≠0 if N≠N'	Compute time	Compare any networks?	Common structure?
rNNI	YES	NP-hard	NO (# retics***)	YES
rSPR / rTBR	YES	$O(2.4^d\, n)$ Linear kernel	NO (# retics)	YES
Cherry-picking	YES	$O(3^r\, n^c)$	NO (orchards)	YES
Contractions	YES	NP-hard	YES, but no Δ-ineq.	YES
d_{LGT}	YES	NP-hard, poly variant	NO (LGT networks)	YES
ominus ominus-SF	YES	NP-hard FPT	YES	YES

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

Clusters in a network

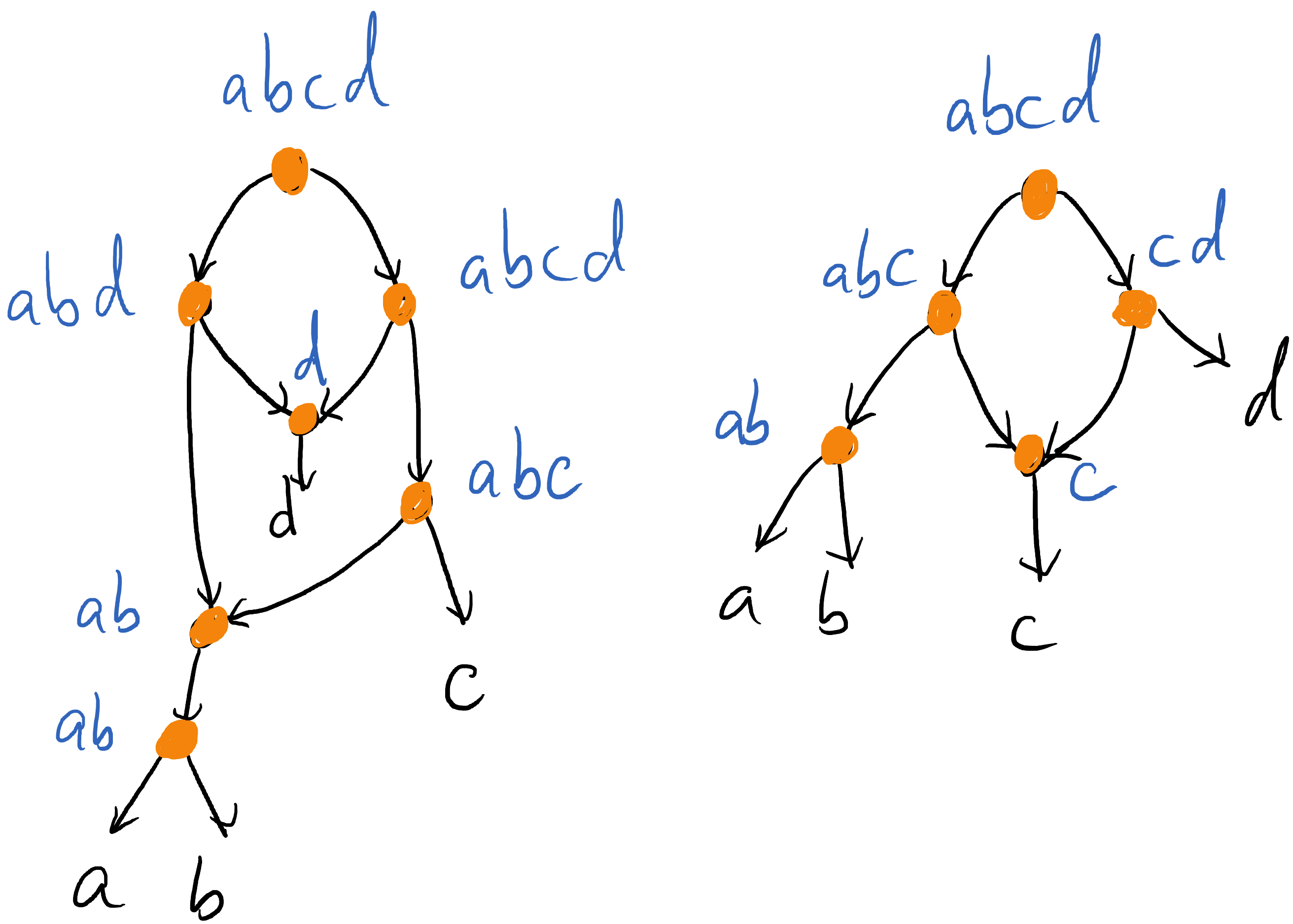
- For a vertex v, $cluster(v)$ = set of leaf taxa that v can reach
- Sometimes called *hardwired clusters*





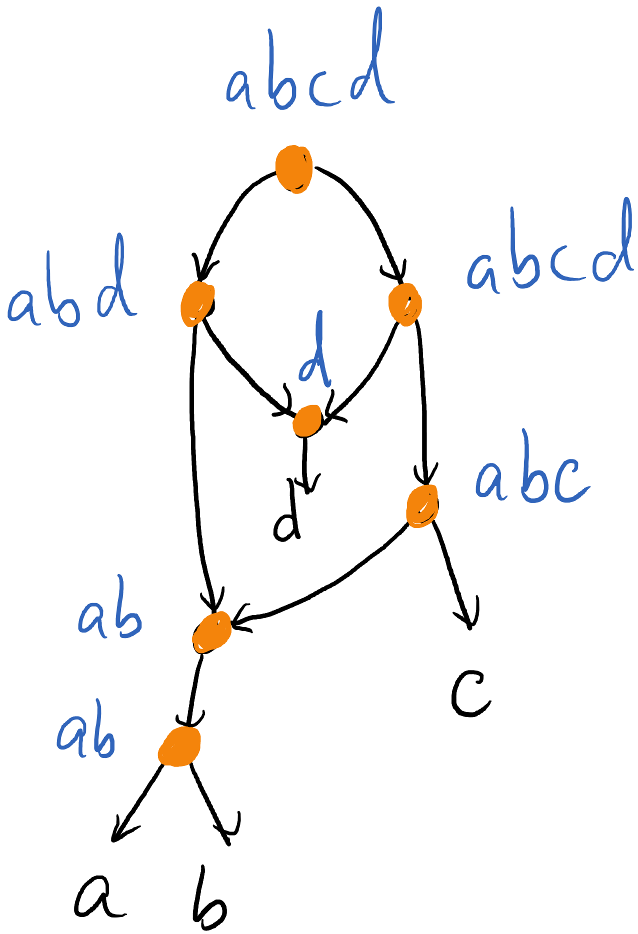
Clusters in a network

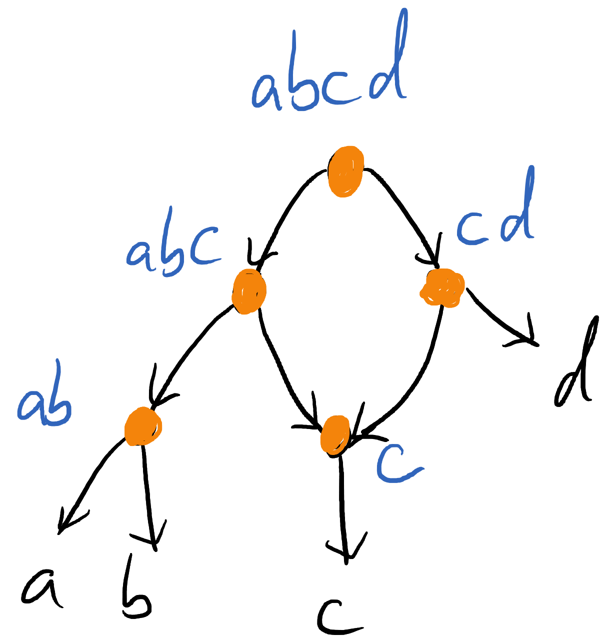
- For a vertex v, $cluster(v)$ = set of leaf taxa that v can reach
- Sometimes called *hardwired clusters*



Hardwired clusters distance

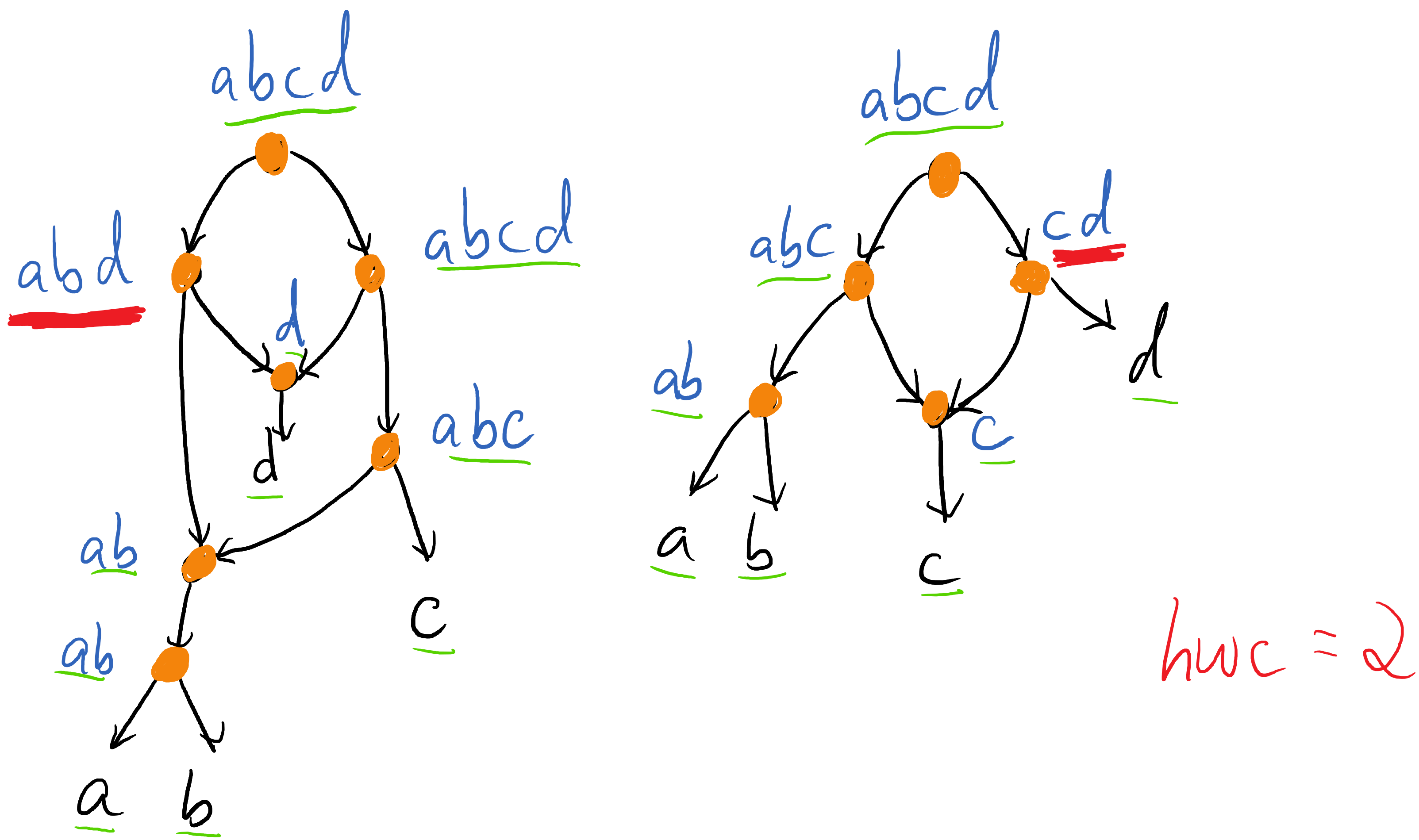
- $hwc(N_1, N_2)$ = number of clusters in one network not in the other.





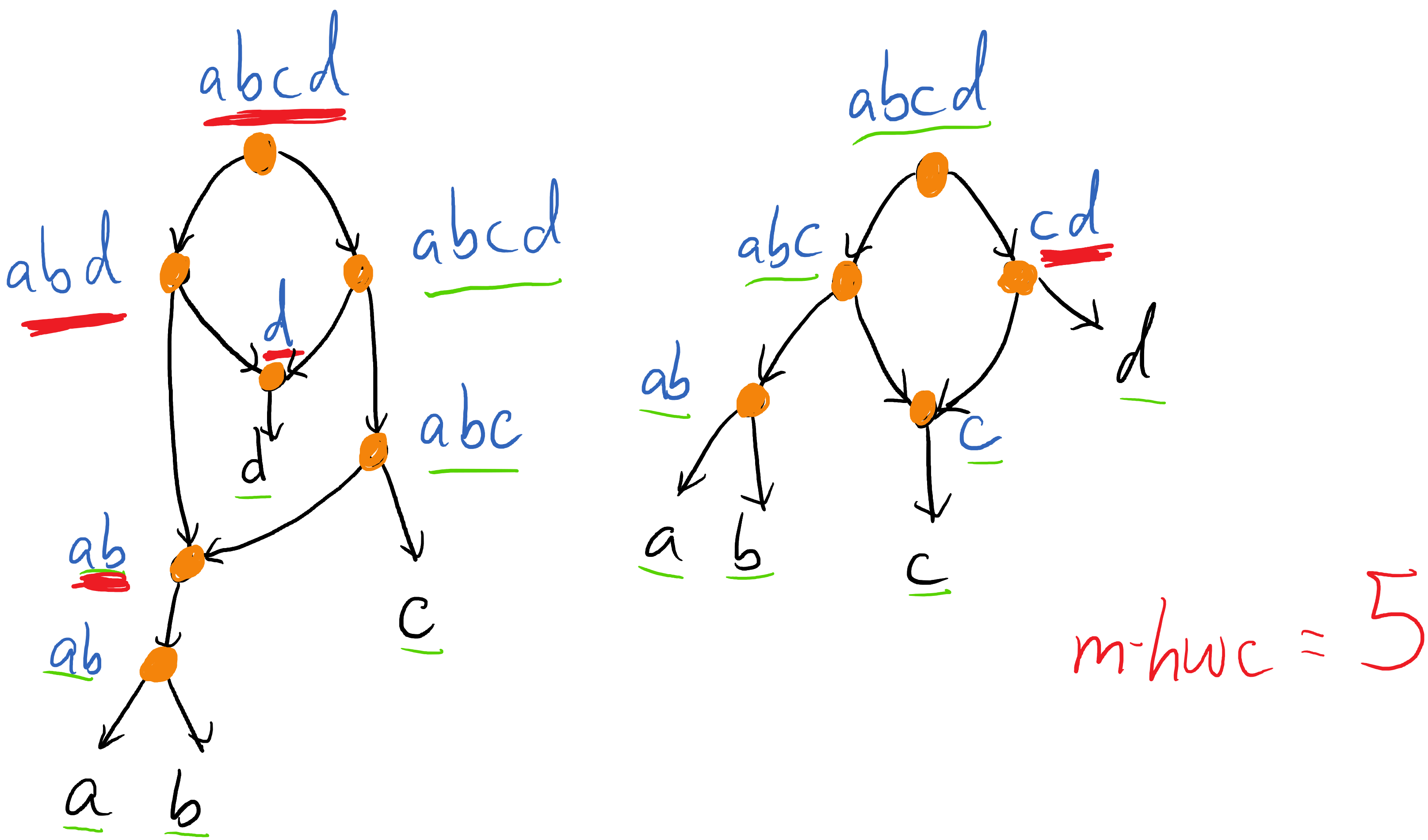
Hardwired clusters distance

- $hwc(N_1, N_2)$ = number of clusters in one network not in the other.



Hardwired clusters distance

- $hwc(N_1, N_2)$ = number of clusters in one network not in the other.
- $m\text{-}hwc(N_1, N_2)$ = same but count differences in multiplicites.

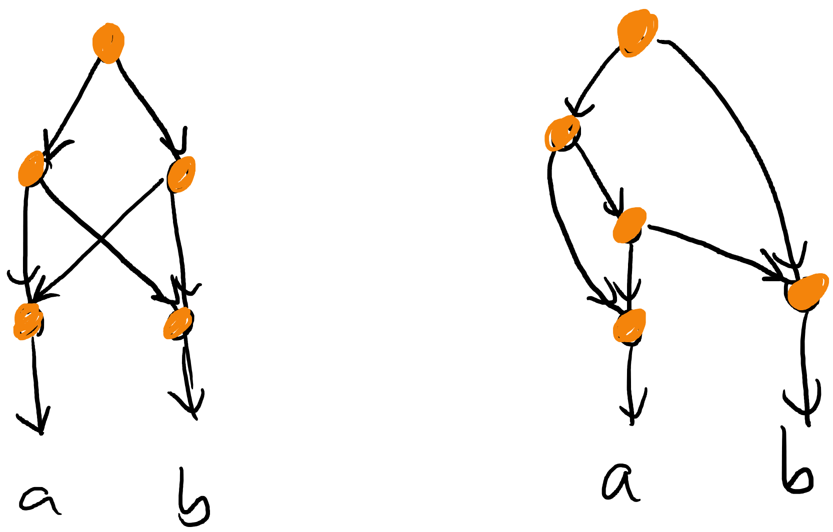


Distinguishability problem

- $hwc(N_1, N_2)$ can be 0 even if N_1 and N_2 are different.
- Can lead to overconfident conclusions!

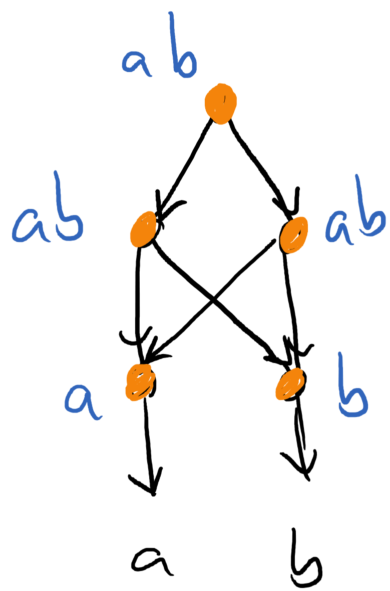
Distinguishability problem

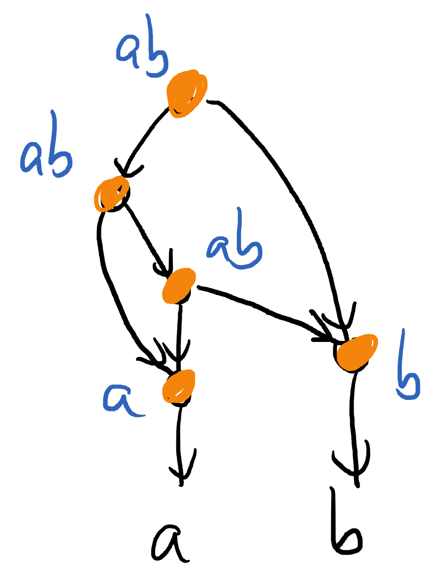
- $hwc(N_1, N_2)$ can be 0 even if N_1 and N_2 are different.
- Can lead to overconfident conclusions!



Distinguishability problem

- $hwc(N_1, N_2)$ can be 0 even if N_1 and N_2 are different.
- Can lead to overconfident conclusions!





Distinguishability problem

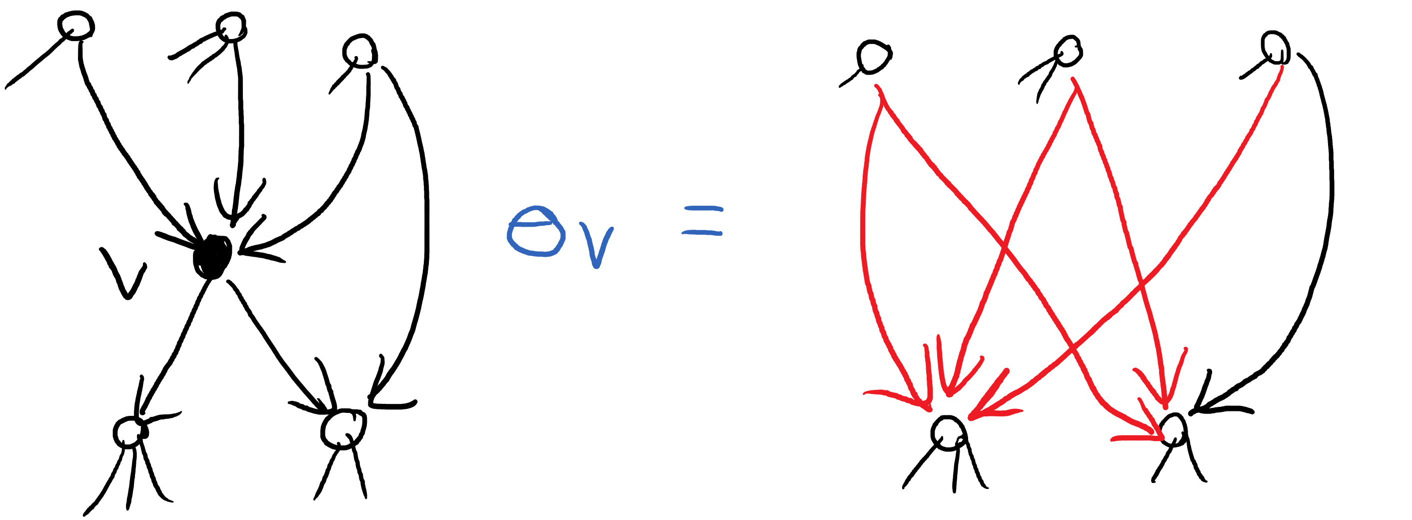
- $hwc(N_1, N_2)$ can be 0 even if N_1 and N_2 are different.
- Can lead to overconfident conclusions!

Solution

- Define a distance based on the **minimum number of clusters to delete** in the networks to **make them identical**.
- What does it mean to delete a cluster?
- $\ominus$-operation

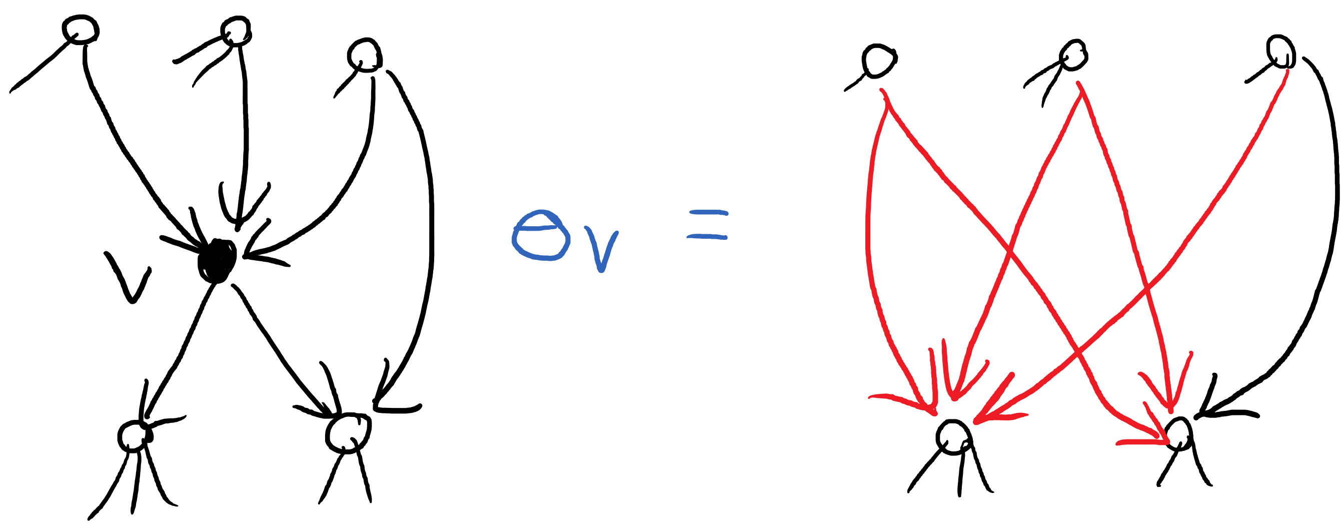
The ominus $\ominus$ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs



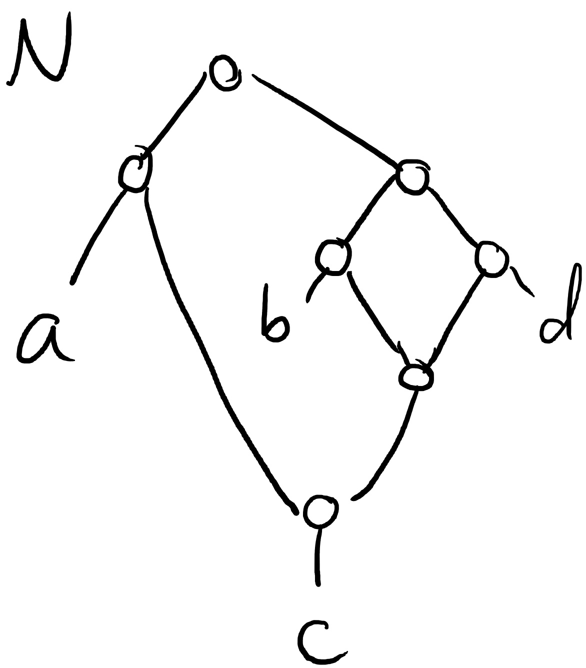
The ominus $\ominus$ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs
- Introduced in [Shanavas, Changat, Hellmuth, Stadler, 2024]
- Eliminates **redundant vertices** that are the LCA of no two leaves.



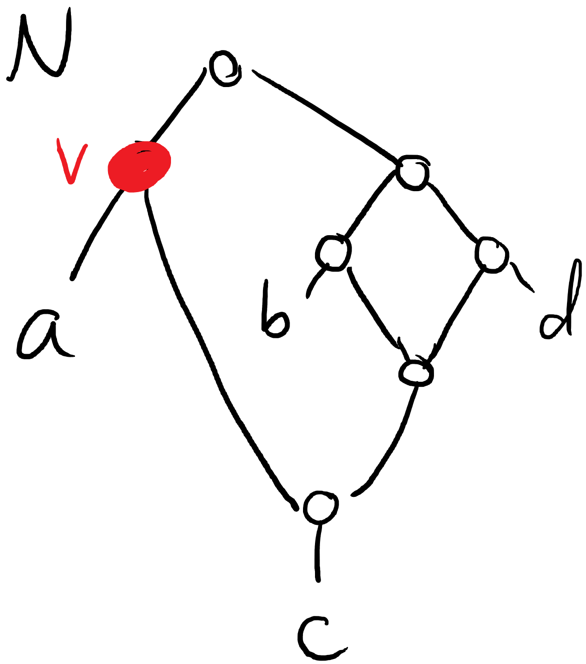
The ominus $\ominus$ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs



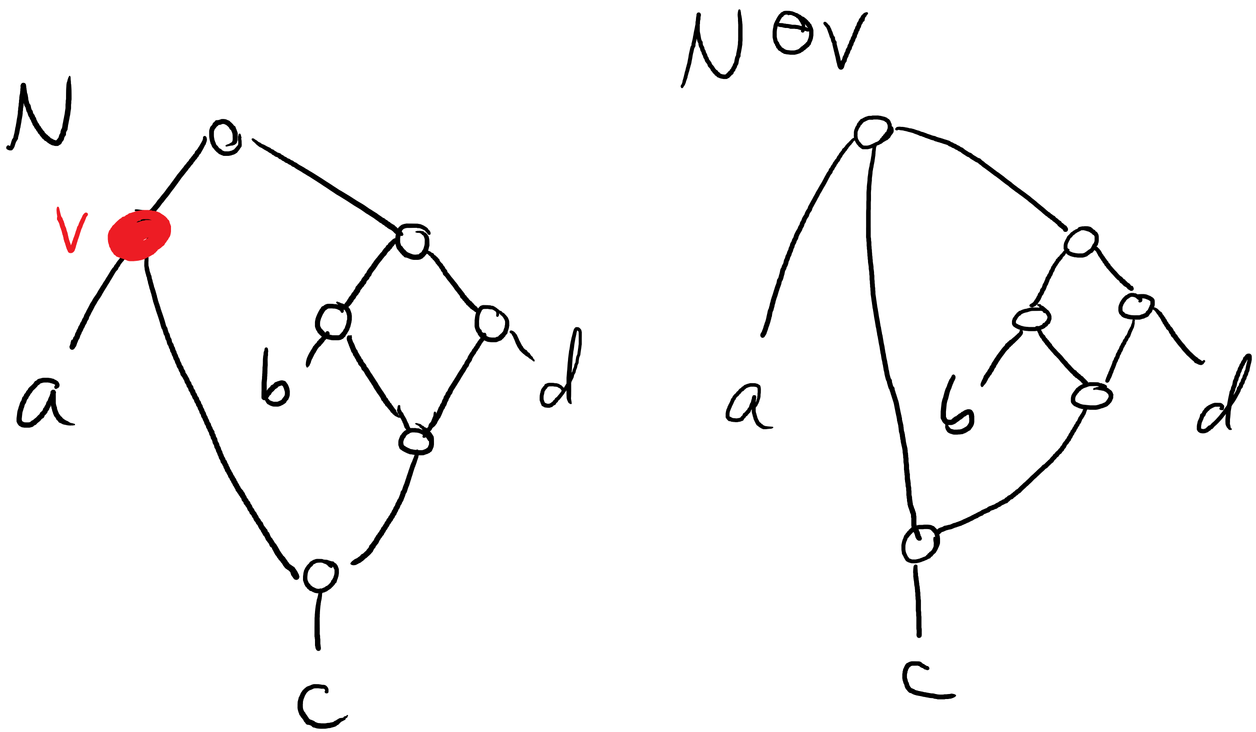
The ominus $\ominus$ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs



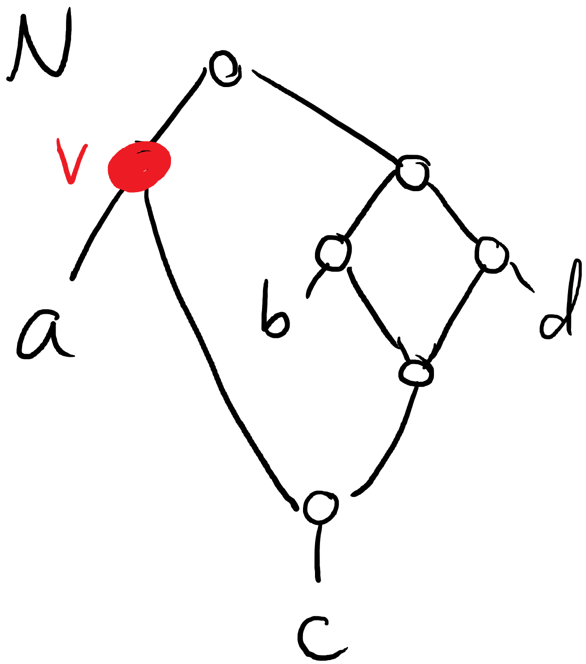
The ominus $\ominus$ operator

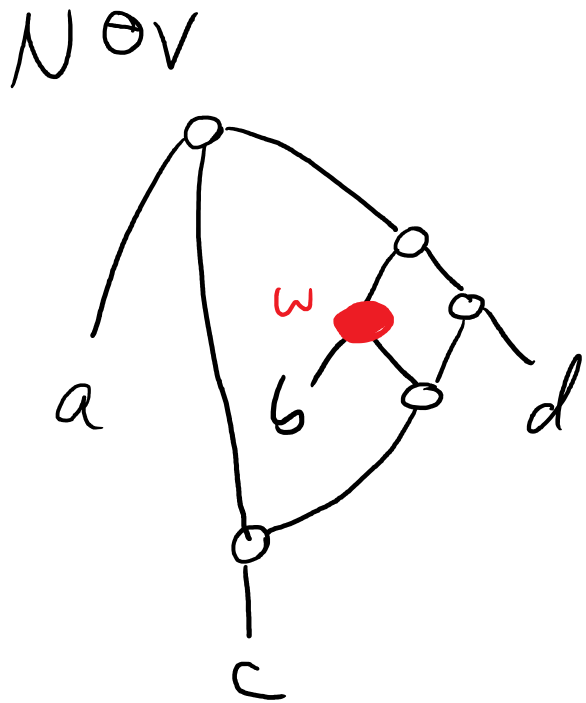
- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs



The ominus $\ominus$ operator

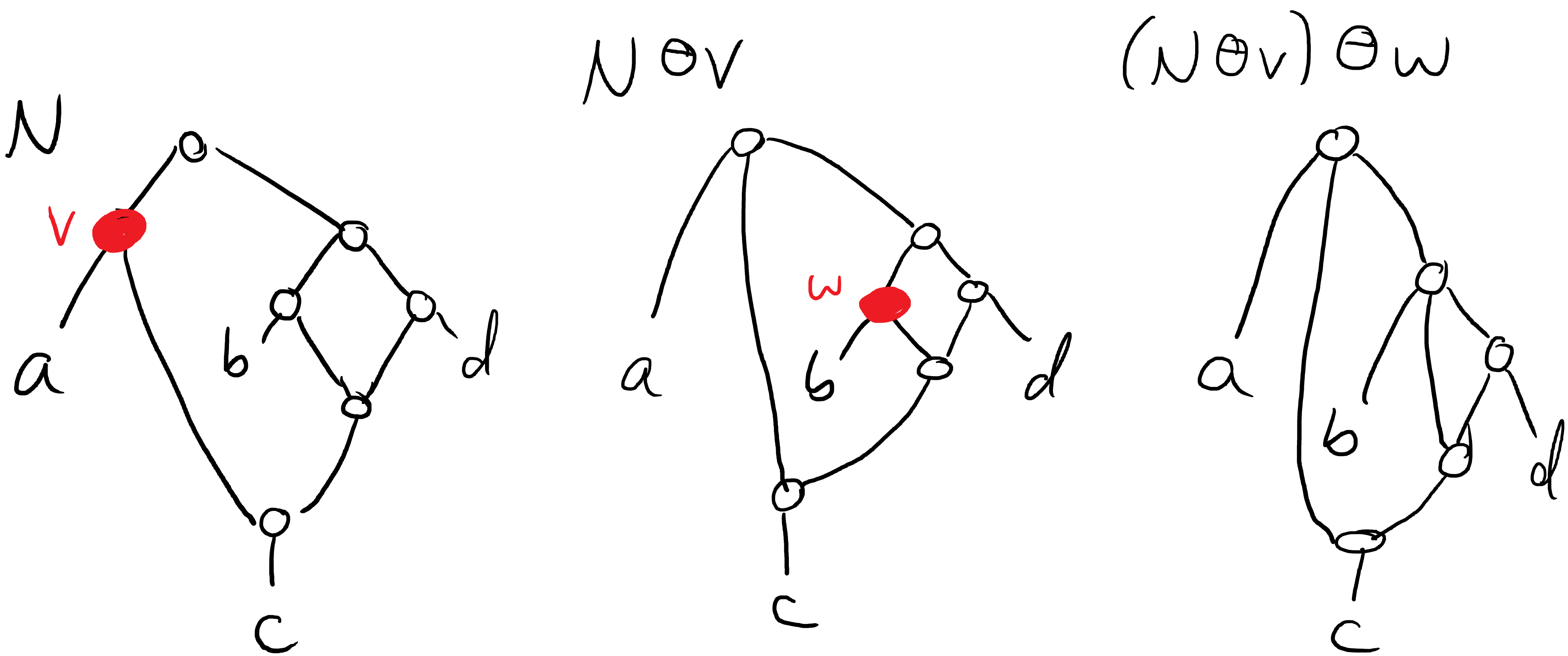
- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs





The ominus ⊖ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs



The ominus $\ominus$ operator

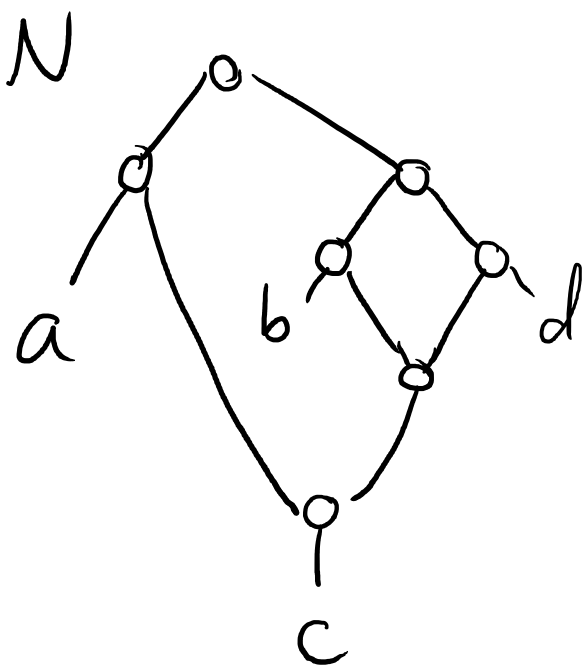
- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs

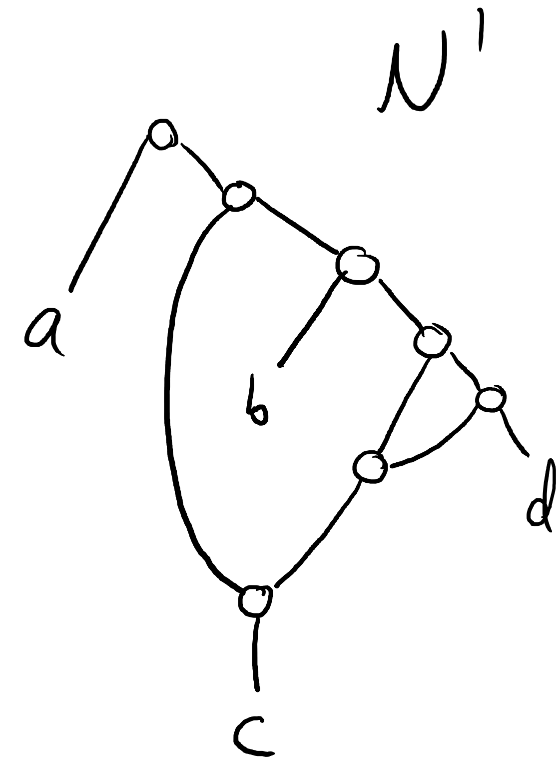
$dist_{\ominus}(N, N') =$ minimum # of $\ominus$ to apply to N and N' to obtain isomorphic networks

The ominus $\ominus$ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs

$dist_{\ominus}(N, N') =$ minimum # of $\ominus$ to apply to N and N' to obtain isomorphic networks

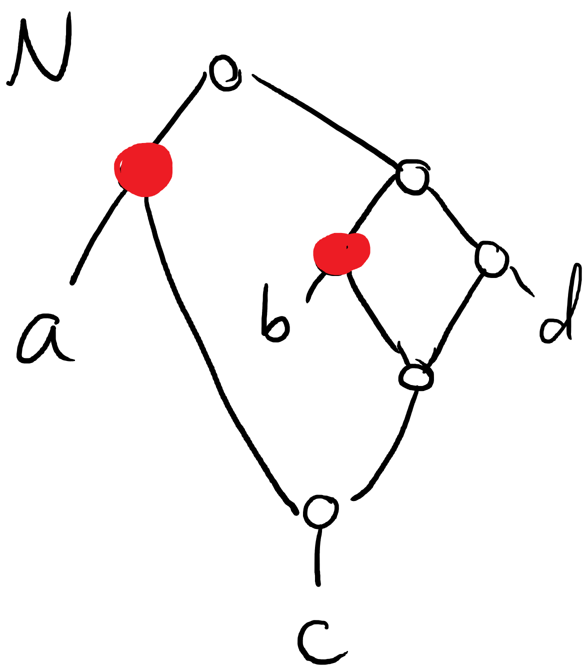


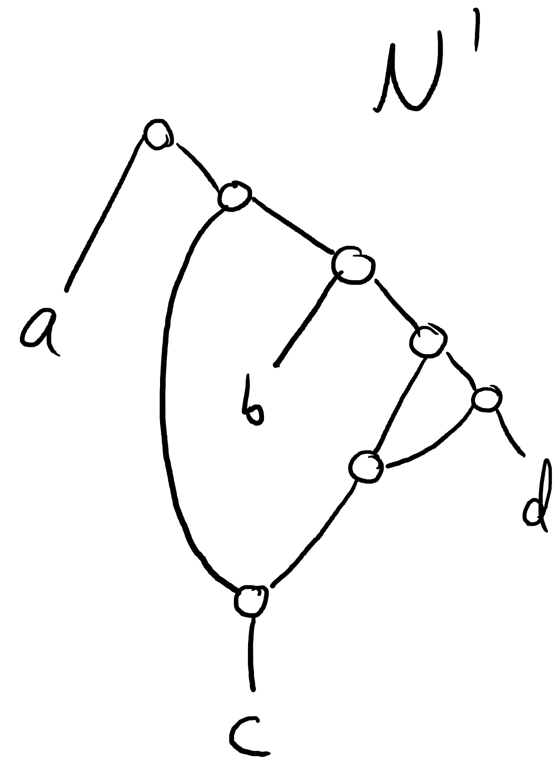


The ominus $\ominus$ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs

$dist_{\ominus}(N, N') =$ minimum # of $\ominus$ to apply to N and N' to obtain isomorphic networks

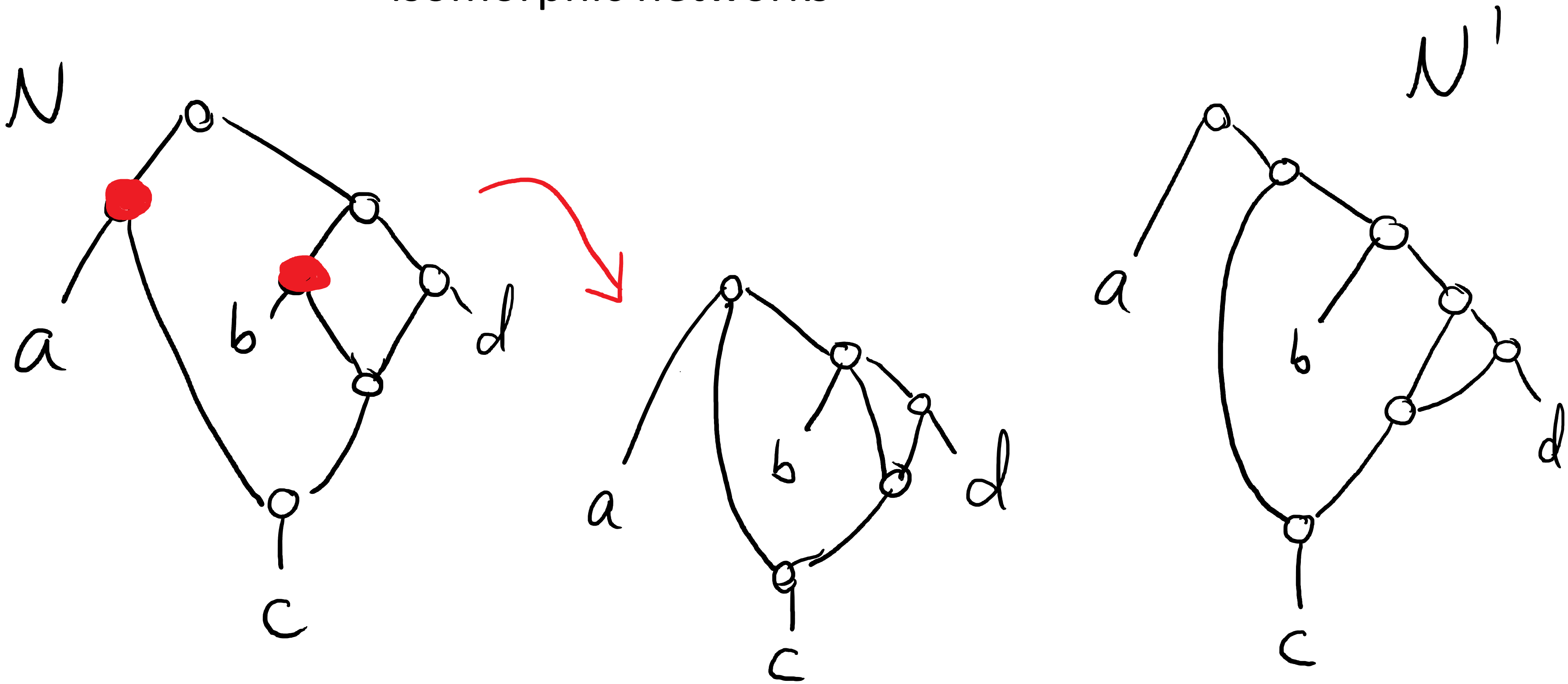




The ominus $\ominus$ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs

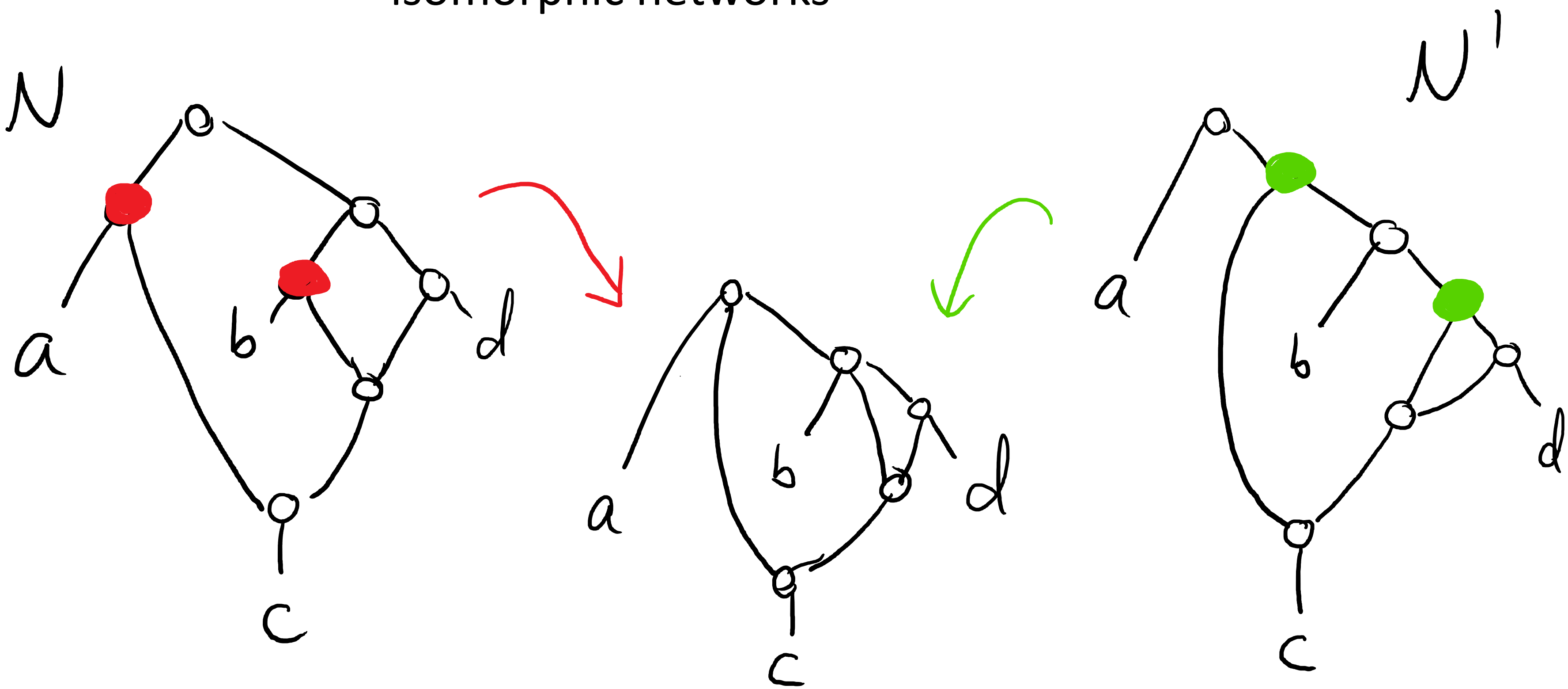
$dist_{\ominus}(N, N') =$ minimum # of $\ominus$ to apply to N and N' to obtain isomorphic networks



The ominus $\ominus$ operator

- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs

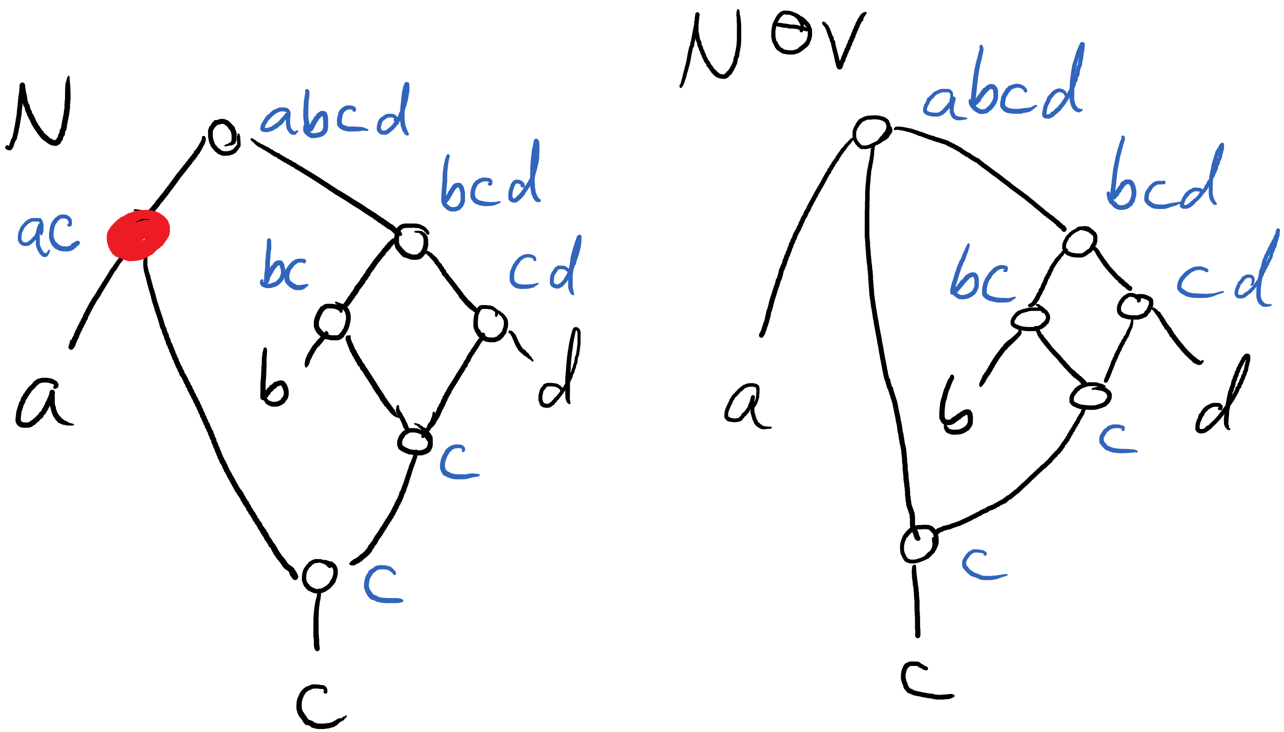
$dist_{\ominus}(N, N') =$ minimum # of $\ominus$ to apply to N and N' to obtain isomorphic networks



Property 1: $\ominus$ deletes exactly one cluster, preserves others

Property 1: $\ominus$ deletes exactly one cluster, preserves others.

i.e., $\textit{multiset-clusters}(N \ominus v) = \textit{multiset-clusters}(N) \setminus \{\textit{cluster}(v)\}$



Property 2: $dist_{\ominus}(N, N') \geq m\text{-}hwc(N, N') \geq hwc(N, N')$

$hwc(N, N') =$ hardwired clusters distance = $|clusters(N) \Delta clusters(N')|$

$m\text{–}hwc(N, N') = |multiset\text{-}clusters(N) \Delta multiset\text{-}clusters(N')|$

Property 2: $dist_{\ominus}(N, N') \geq m\text{−}hwc(N, N') \geq hwc(N, N')$

$hwc(N, N')$ = hardwired clusters distance = $|clusters(N) \Delta clusters(N')|$

$m\text{−}hwc(N, N') = |multiset\text{-}clusters(N) \Delta multiset\text{-}clusters(N')|$

Both hwc and $m\text{−}hwc$ can be 0 on distinct networks.

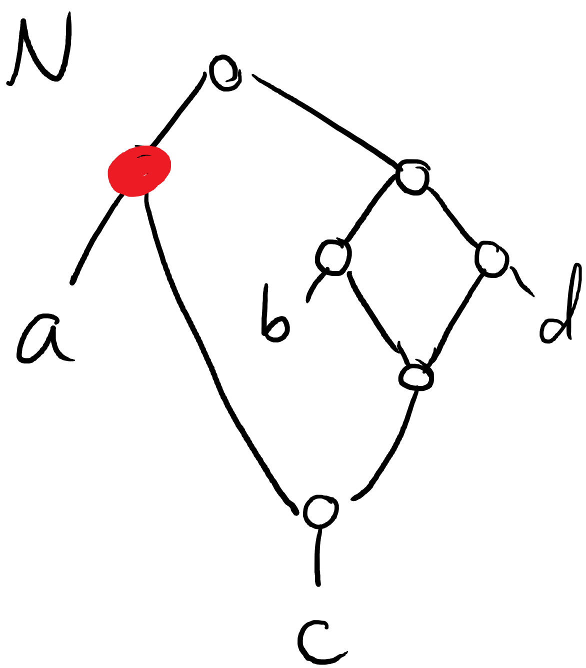
Not $dist_{\ominus}$:)

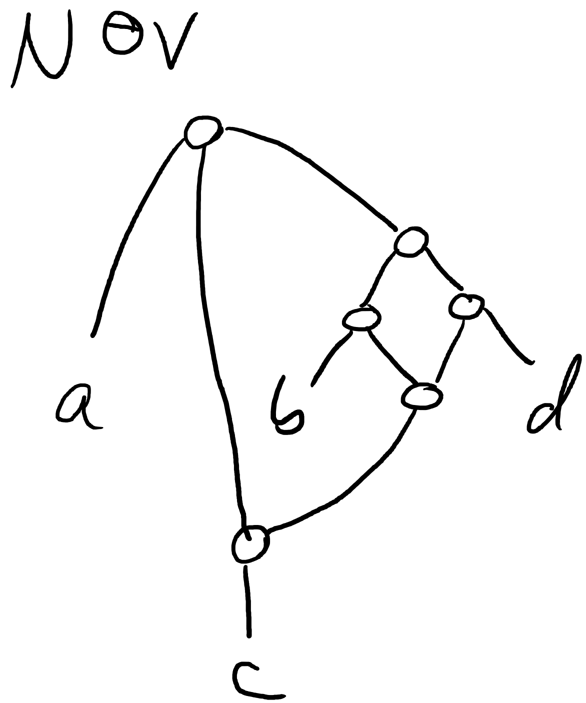
In $dist_{\ominus}$, clusters in $clusters(N) \Delta clusters(N')$ must be $\ominus$'d. But does not stop there if not enough.

Property 3: $\ominus$ preserves ancestry relations

For $u, w \in V(N \ominus v)$, $u \preccurlyeq_{N \ominus v} w$ if and only if $u \preccurlyeq_N w$.

$\preccurlyeq$ means "descends from"



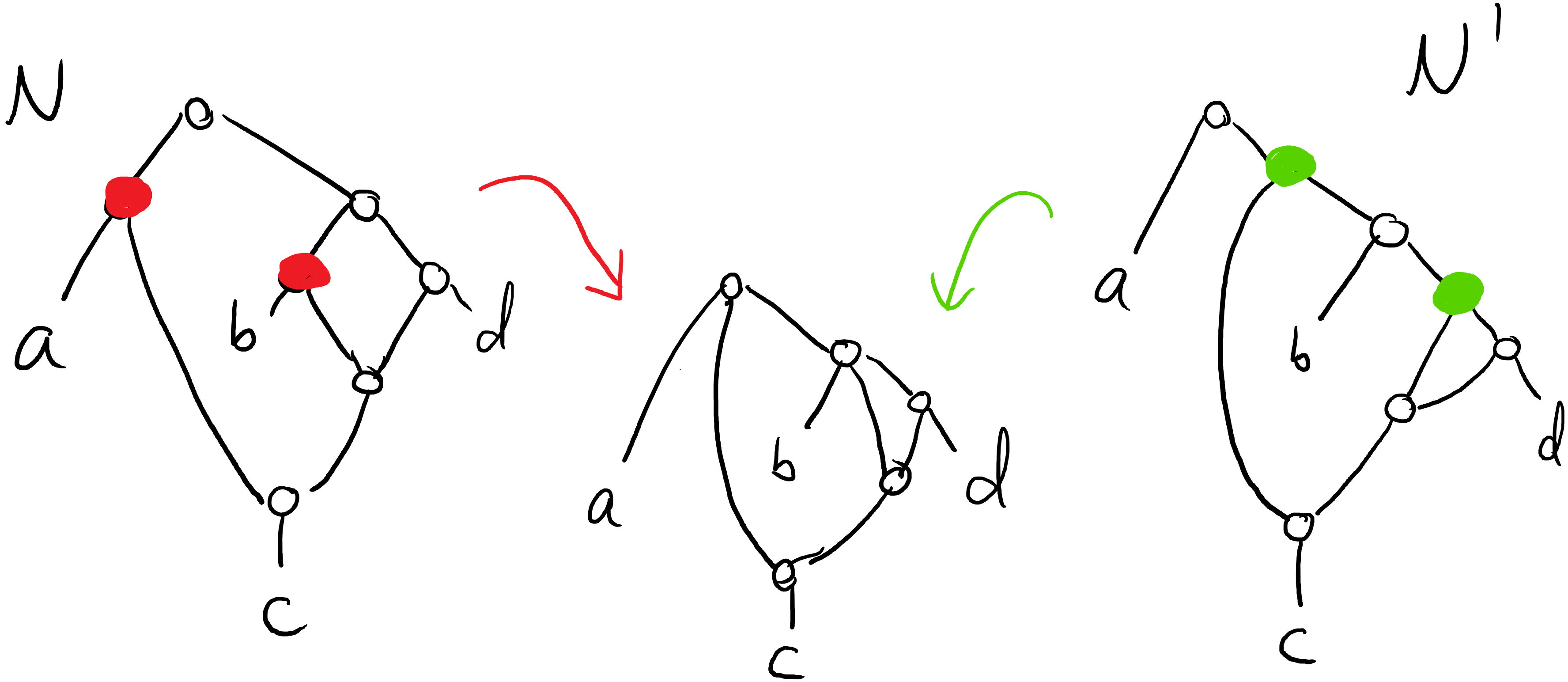


Property 3: $\ominus$ preserves ancestry relations

For $u, w \in V(N \ominus v)$, $u \preccurlyeq_{N \ominus v} w$ if and only if $u \preccurlyeq_N w$.

$\preccurlyeq$ means "descends from"

Property 4: $dist_{\ominus}(N, N')$ gives, as a by-product, a maximum common structure that **agrees on clusters and on ancestry relations**.



Property 5: $dist_{\ominus}(N, N')$ is a **metric** on the space of **all networks** with the same leaves.

1. $dist_{\ominus}(N,N')$ is well-defined for all N, N' on X.
2. $dist_{\ominus}(N, N') = 0 \Leftrightarrow N \simeq N'$.
3. $dist_{\ominus}(N, N') = d(N', N)$.
4. $dist_{\ominus}(N, N') \leq d(N, N'') + d(N'', N')$.

Property 5: $dist_{\ominus}(N, N')$ is a **metric** on the space of **all networks** with the same leaves.

1. $dist_{\ominus}(N,N')$ is well-defined for all N, N' on X.

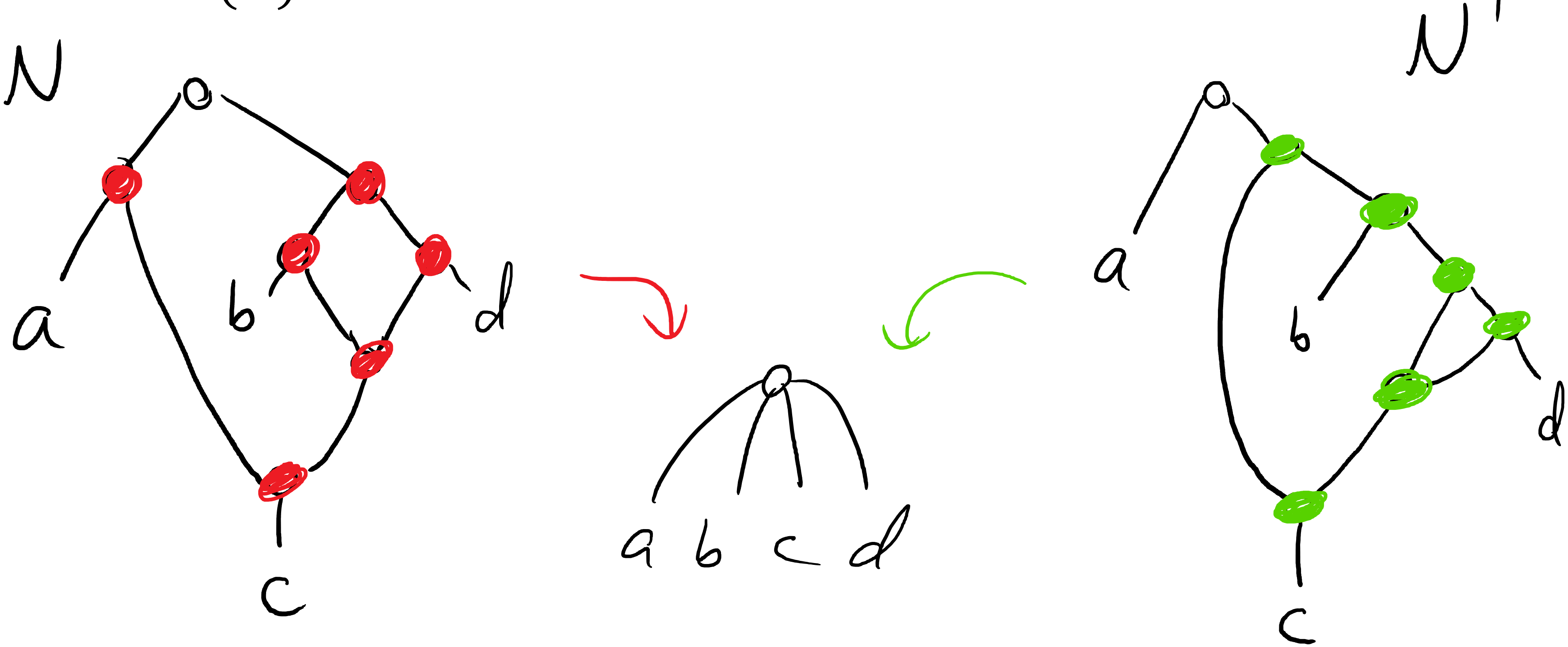
Property 5: $dist_{\ominus}(N, N')$ is a **metric** on the space of **all networks** with the same leaves.

1. $dist_{\ominus}(N,N')$ is well-defined for all N, N' on X.

Because any network can be reduced to a star tree.

$$0 \leq dist_{\ominus}(N, N') \leq |V^0(N)| + |V^0(N')|$$

where $V^0(N)$ is the set of internal vertices.



Property 5: $dist_{\ominus}(N, N')$ is a **metric** on the space of **all networks** with the same leaves.

1. $dist_{\ominus}(N,N')$ is well-defined for all N, N' on X.

Because any network can be reduced to a star tree.

$$0 \leq dist_{\ominus}(N, N') \leq |V^0(N)| + |V^0(N')|$$

where $V^0(N)$ is the set of internal vertices.

In fact, $|V^0(N)| + |V^0(N')|$ is the diameter.

Property 5: $dist_{\ominus}(N, N')$ is a **metric** on the space of **all networks** with the same leaves.

1. $dist_{\ominus}(N,N')$ is well-defined for all N, N' on X.
2. $dist_{\ominus}(N, N') = 0 \Leftrightarrow N \simeq N'$.
3. $dist_{\ominus}(N, N') = d(N', N)$.
4. $dist_{\ominus}(N, N') \leq d(N, N'') + d(N'', N')$.

Property 5: $dist_{\ominus}(N, N')$ is a **metric** on the space of **all networks** with the same leaves.

1. $dist_{\ominus}(N,N')$ is well-defined for all N, N' on X.
2. $dist_{\ominus}(N, N') = 0 \Leftrightarrow N \simeq N'$. **By definition.**
3. $dist_{\ominus}(N, N') = d(N', N)$. **By definition.**
4. $dist_{\ominus}(N, N') \leq d(N, N'') + d(N'', N')$. **Needs argument.**

Theorem: Computing $dist_{\ominus}(N, N')$ is NP-complete, even on **distinct-cluster networks**.

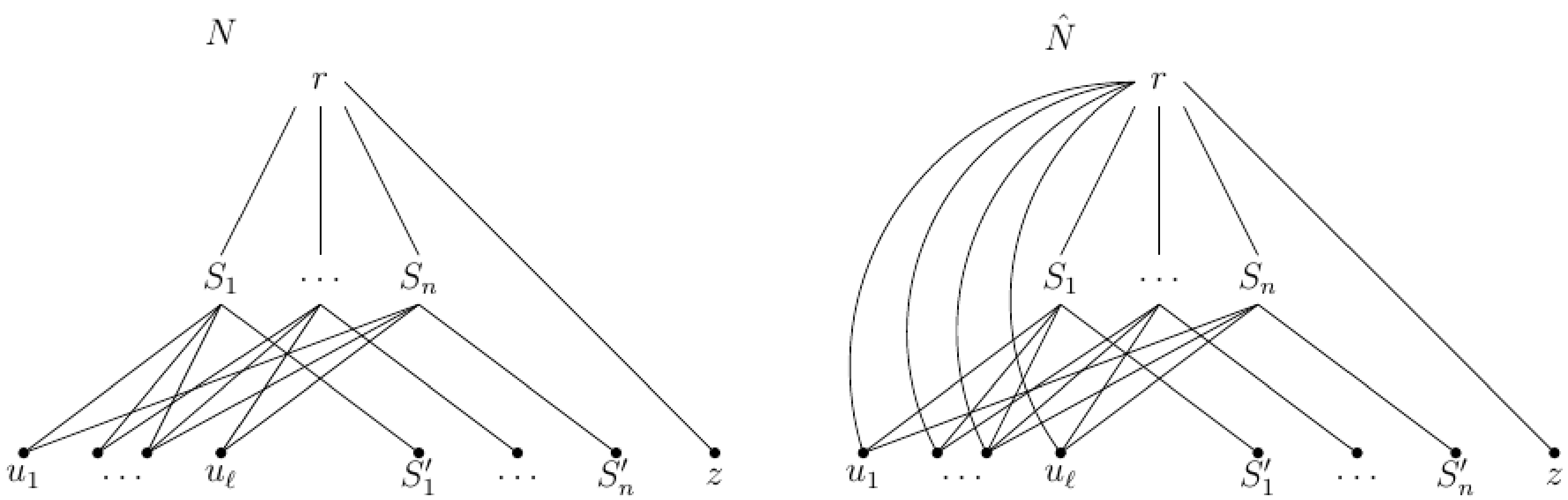
(In fact, it is SET-COVER-hard, so (probably) no FPT algorithm, no constant-factor approx.)

Distinct-cluster means that each cluster has multiplicity 1.

Theorem: Computing $dist_{\ominus}(N, N')$ is NP-complete, even on **distinct-cluster networks**.

(In fact, it is SET-COVER-hard, so (probably) no FPT algorithm, no constant-factor approx.)

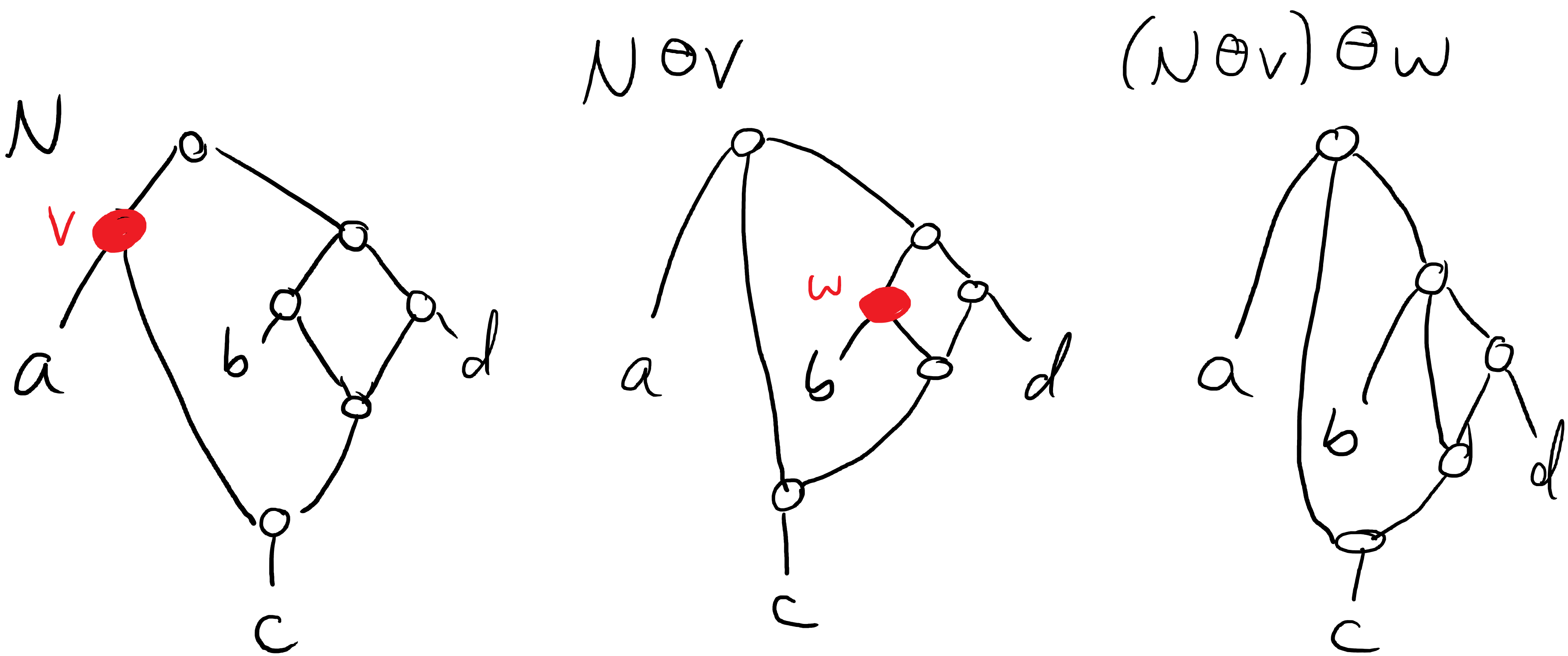
Distinct-cluster means that each cluster has multiplicity 1.



Property 6: $\ominus$ can create new shortcuts

Shortcut = arc (u, v) such that there is a $u - v$ path without that arc.

Not so desirable for several reasons, including computational complexity.

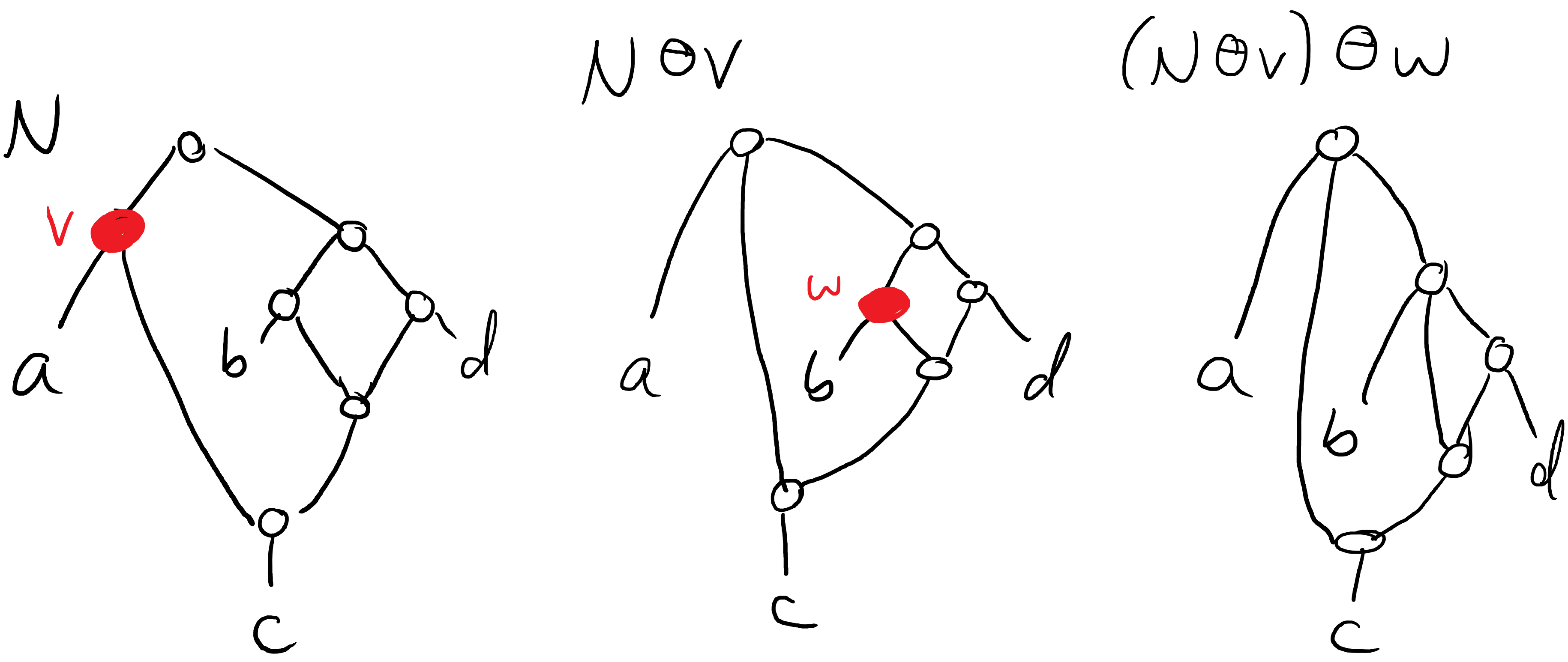


Property 6: $\ominus$ can create new shortcuts

Shortcut = arc (u, v) such that there is a $u - v$ path without that arc.

Not so desirable for several reasons, including computational complexity.

Define $\ominus^-$: like $\ominus$, but delete shortcuts whenever they appear.



The shortcut-free $\ominus$-distance

$$dist_{\ominus}^{-}(N, N') = min\{|W| + |W'| : N \ominus W \simeq_{SF} N' \ominus W'\}$$

where $\simeq_{SF}$ means "isomorphic after removing shortcuts".

The shortcut-free $\ominus$-distance

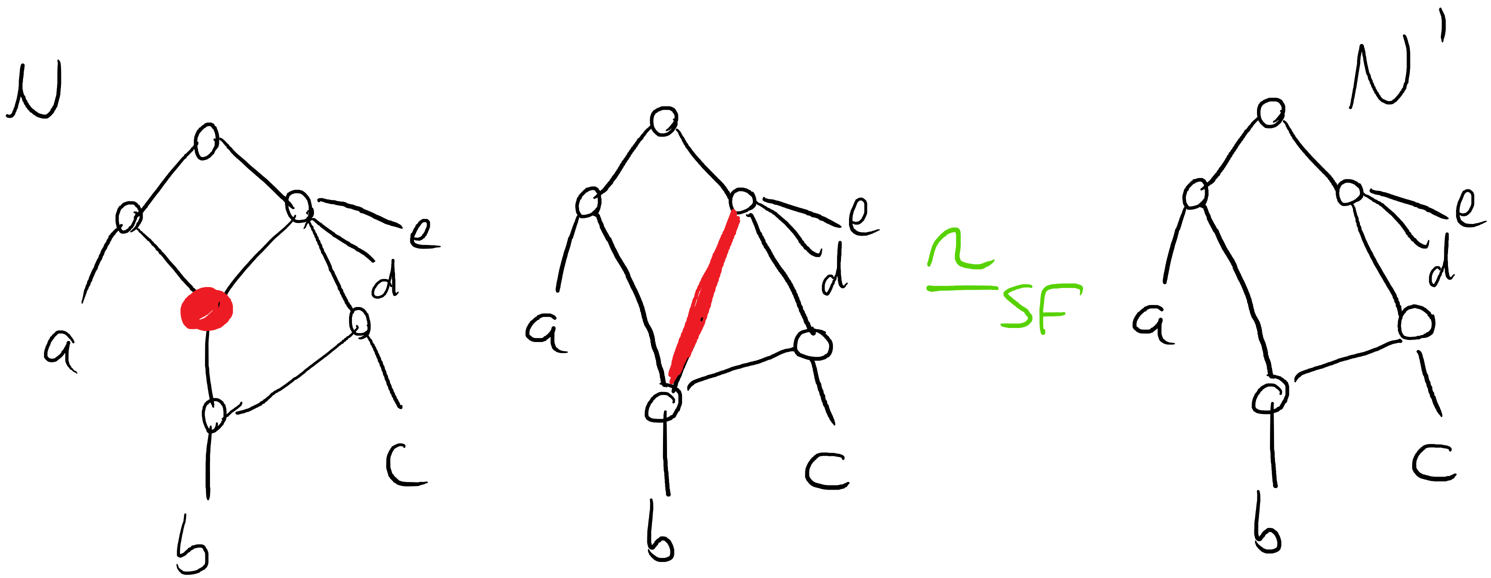
$$dist^{-}_{\ominus}(N, N') = min\{|W| + |W'| : N \ominus W \simeq_{SF} N' \ominus W'\}$$

where $\simeq_{SF}$ means "isomorphic after removing shortcuts".

The shortcut-free ⊖-distance

$$dist_{\ominus}^{-}(N, N') = min\{|W| + |W'| : N \ominus W \simeq_{SF} N' \ominus W'\}$$

where $\simeq_{SF}$ means "isomorphic after removing shortcuts".



The shortcut-free $\ominus$-distance

$$dist_{\ominus}^{-}(N,N') = min\{|W| + |W'| : N \ominus W \simeq_{SF} N' \ominus W'\}$$

All previous properties maintained.

- $dist_{\ominus}^{-}(N,N') \geq m\text{–}hwc(N,N') \geq hwc(N,N')$
- $dist_{\ominus}^{-}(N,N')$ gives a maximum common structure that agrees on clusters and on ancestry relations.
- $dist_{\ominus}^{-}$ is a metric on the space of all networks on X, with respect to $\simeq_{SF}$
  1. $dist_{\ominus}^{-}(N,N')$ is well-defined for all N, N' on X.
  2. $dist_{\ominus}^{-}(N,N') = 0 \Leftrightarrow N \simeq_{SF} N'$.
  3. $dist_{\ominus}^{-}(N,N') = d(N',N)$.
  4. $dist_{\ominus}^{-}(N,N) \leq d(N,N'') + d(N'',N')$

Theorem: $dist_{\ominus}^{-}(N, N') = m\text{-}hwc(N, N')$ on all the following classes:

1. The class of networks satisfying (PCC)*
2. The class of rooted phylogenetic trees.
3. The class of phylogenetic level-1 networks
4. The class of binary level-1 networks.
5. The class of tree-child networks.
6. The class of normal networks.
7. The class of semi-regular networks;
8. The class of regular networks.

* PCC means $cluster(v) \subseteq cluster(w) \Rightarrow v, w$ are comparable.

Theorem: $dist_{\ominus}^{-}(N, N') = m\text{–}hwc(N, N')$ on all the following classes:

1. The class of networks satisfying (PCC)*
2. The class of rooted phylogenetic trees.
3. The class of phylogenetic level-1 networks
4. The class of binary level-1 networks.
5. The class of tree-child networks.
6. The class of normal networks.
7. The class of semi-regular networks;
8. The class of regular networks.

Theorem: for N, N' distinct-cluster networks, computing $dist_{\ominus}^{-}(N, N')$ is NP-hard.

Theorem: on distinct cluster networks, computing $dist_{\ominus}^{-}$ reduces to the VERTEX-COVER problem (it is twice the vertex-cover number of some graph we call the *bad ancestry* graph).

Corollary: for N, N' distinct-cluster networks, $dist_{\ominus}^{-}(N, N')$ can be computed in time $O(1.12^k poly(n))$, where k is the distance. Also, $dist_{\ominus}^{-}(N, N')$ admits a 2-approximation.

* PCC means $cluster(v) \subseteq cluster(w) \Rightarrow v, w$ are comparable.

Open problems

- **Main problem**: design the perfect metric between networks.
- Our algorithms apply to distinct-cluster networks. How to go beyond?
  - Is $dist_{\ominus}^{-}$ still computable if each cluster has multiplicity at most c?
- Any class for which $dist_{\ominus}$ is tractable?
- Implement and evaluate computation of $dist_{\ominus}$ and $dist_{\ominus}^{-}$