The $\ominus$-metric to compare phylogenetic networks

Marc Hellmuth

Manuel Lafond

Guillaume Scholz

The $\ominus$-metric to compare phylogenetic networks

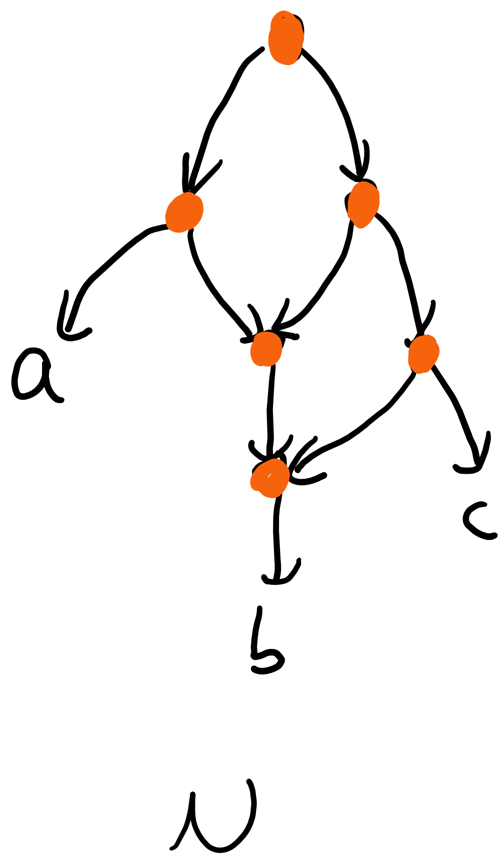
Marc Hellmuth

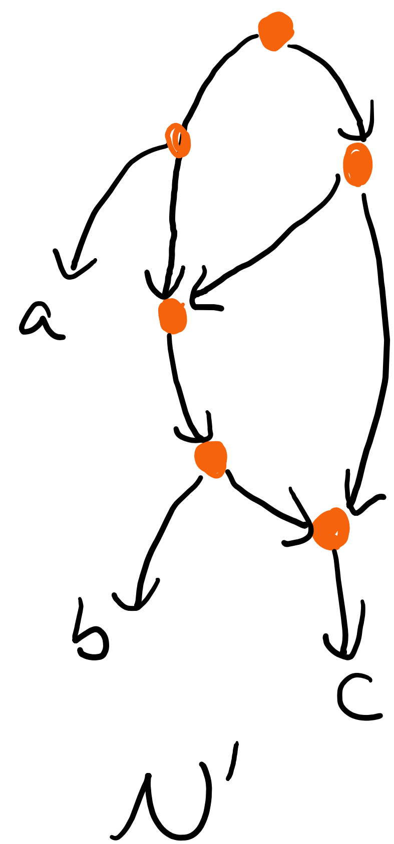
Manuel Lafond

Guillaume Scholz

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 or simulations
- Extract common structure.
- Get a consensus.
- Cluster similar networks.

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

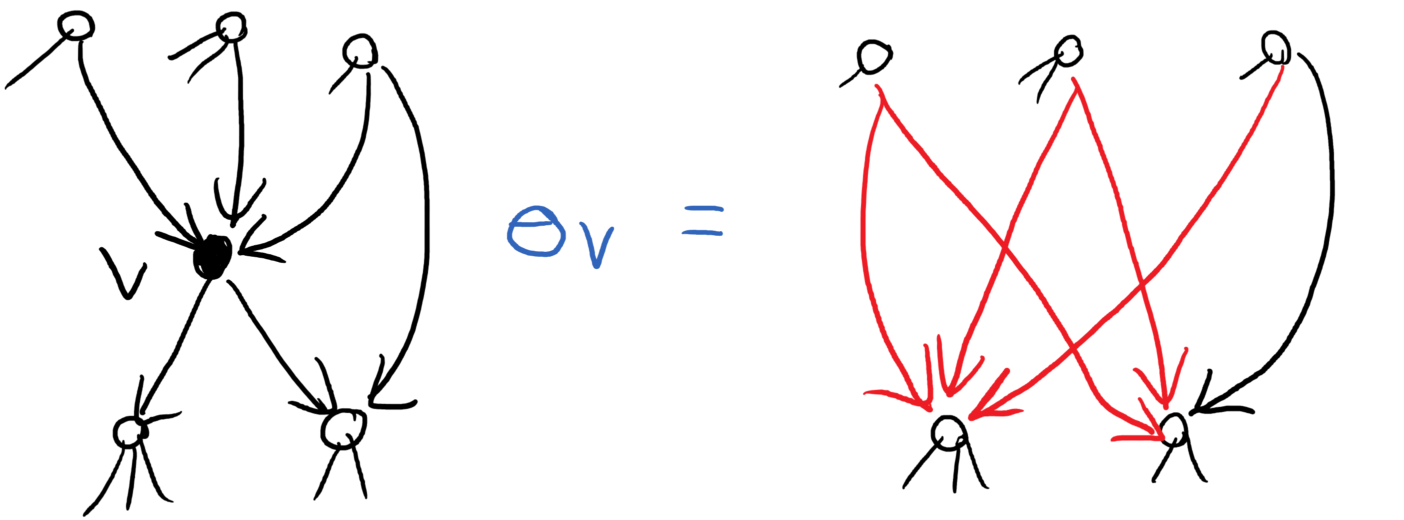
* yes on orchards ** yes on time-consistent tree-child *** add triangle creation move

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

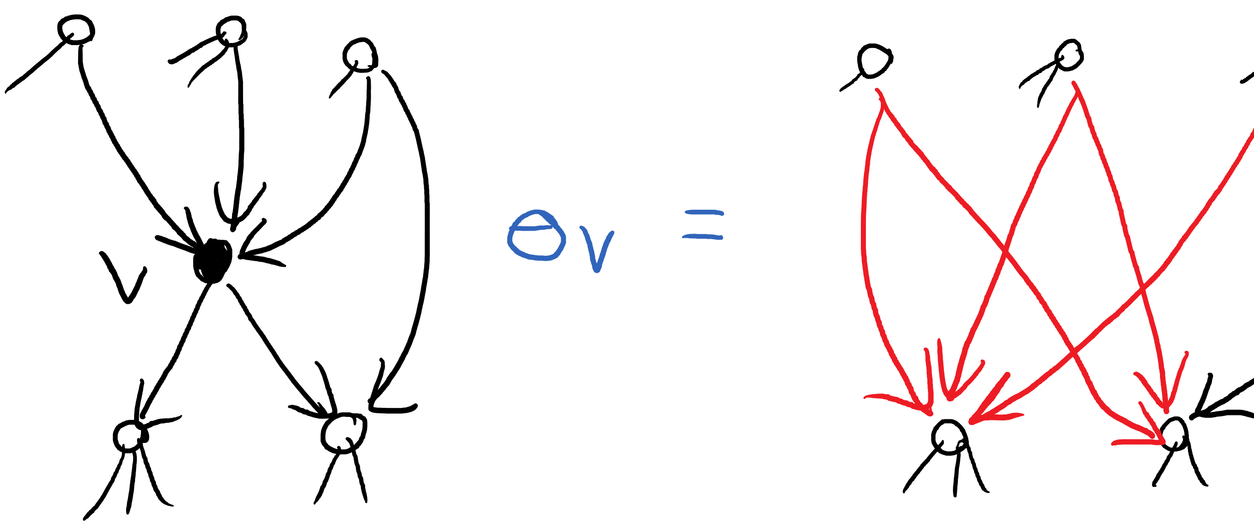
The ominus $\ominus$ operator

- 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



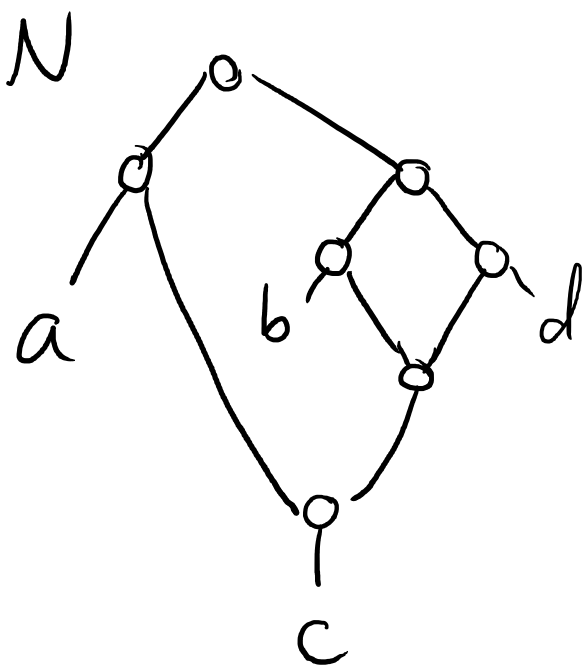
The ominus $\ominus$ operator

- 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.



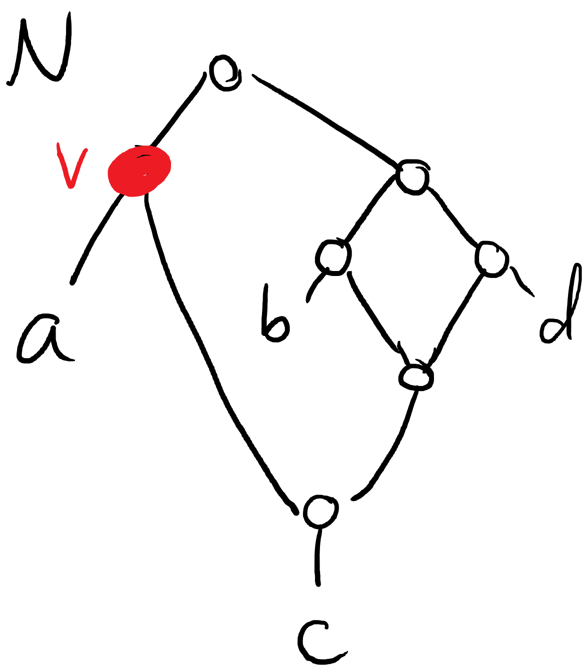
The ominus $\ominus$ operator

- 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



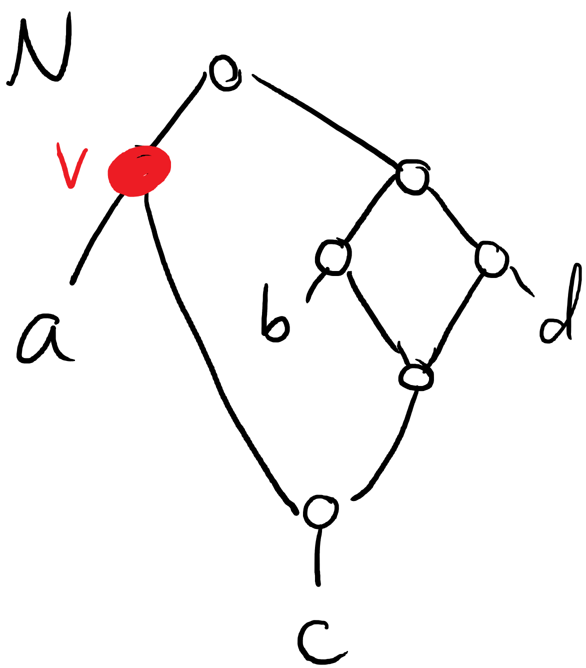
The ominus $\ominus$ operator

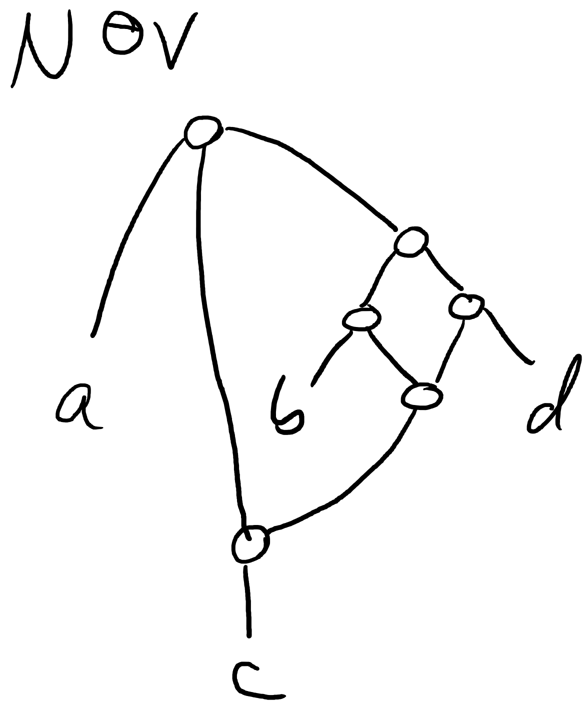
- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs



The ominus $\ominus$ operator

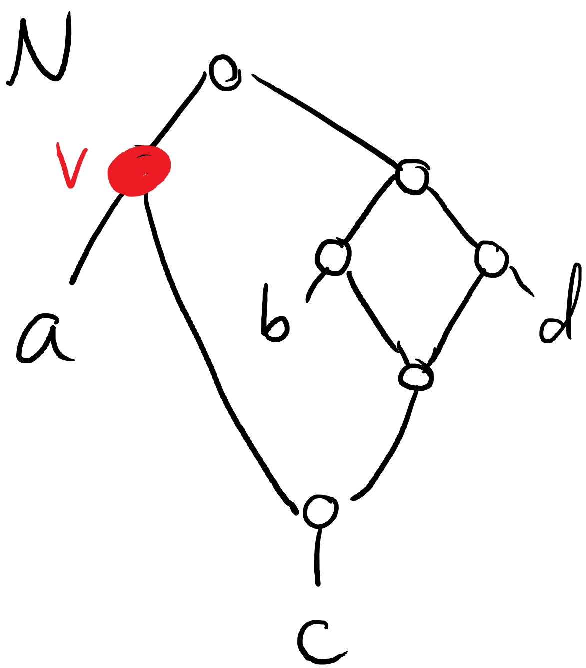
- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs

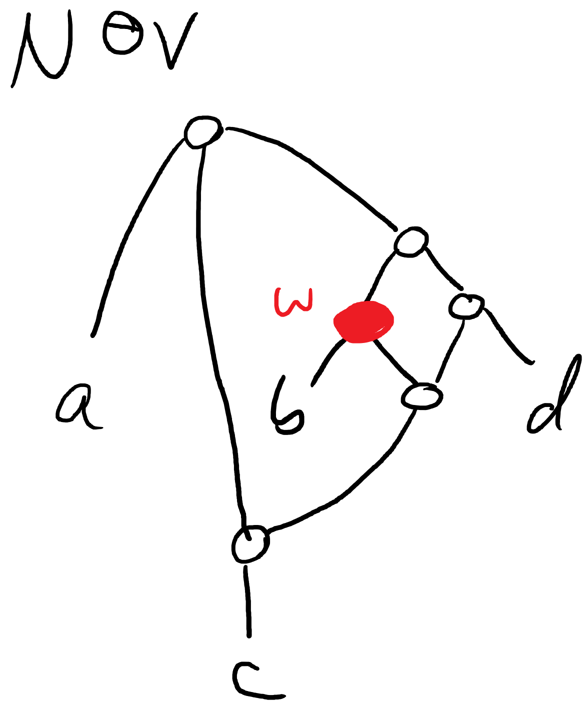




The ominus $\ominus$ operator

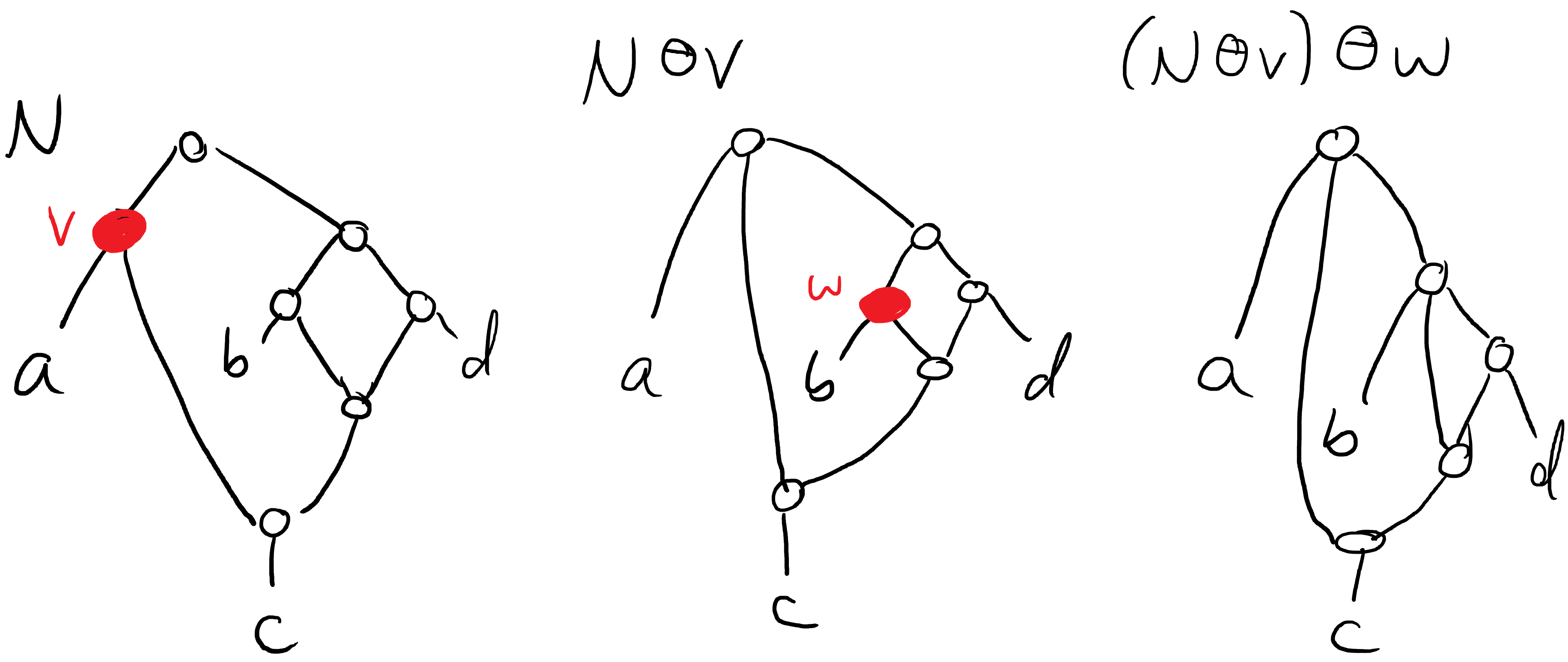
- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs





The ominus $\ominus$ 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

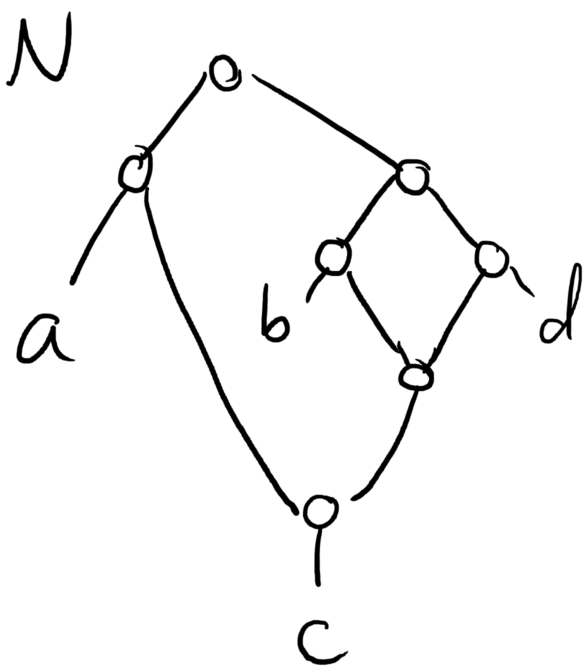
- Choose an internal vertex v of a network N,
- To obtain $N \ominus v$:
  - add all arcs from $parents(v)$ to $children(v)$ (no repeated arcs)
  - delete v and its incident arcs

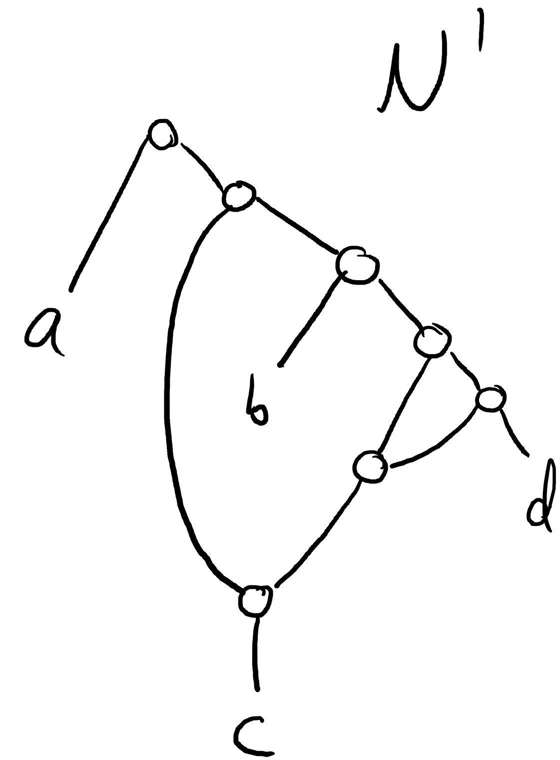
$dist_{\ominus}(N, N') =$ minimum # of $\ominus$ to apply to N and N' to obtain isomorphic networks

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

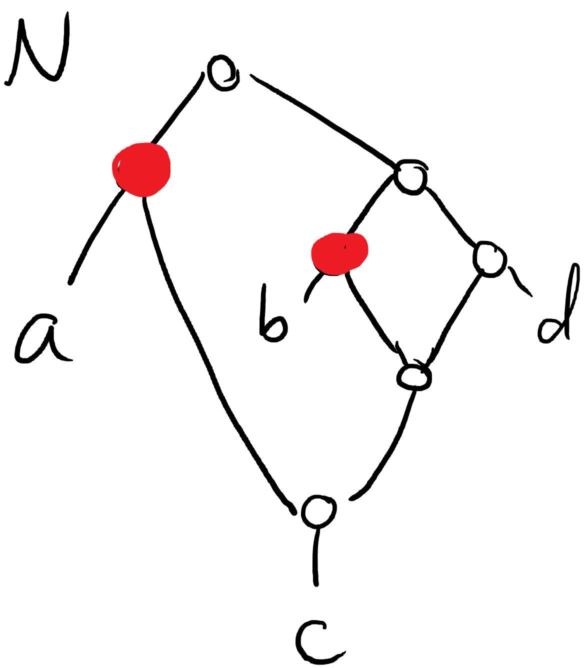


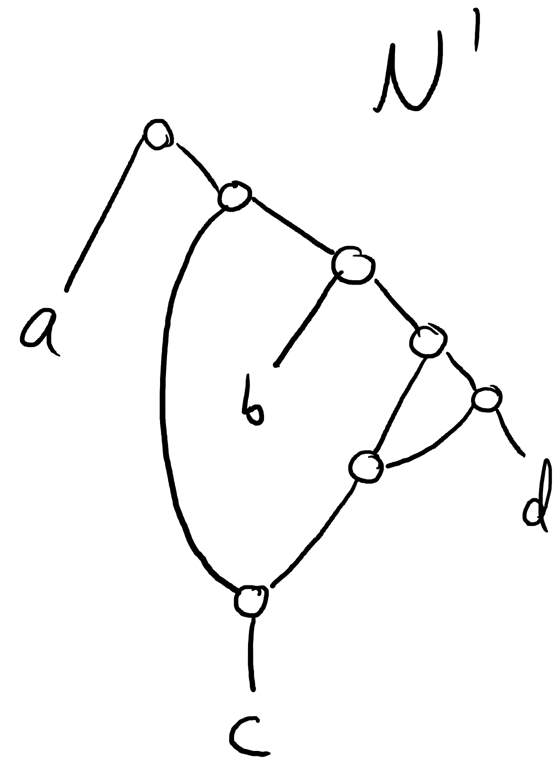


The ominus $\ominus$ operator

- 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

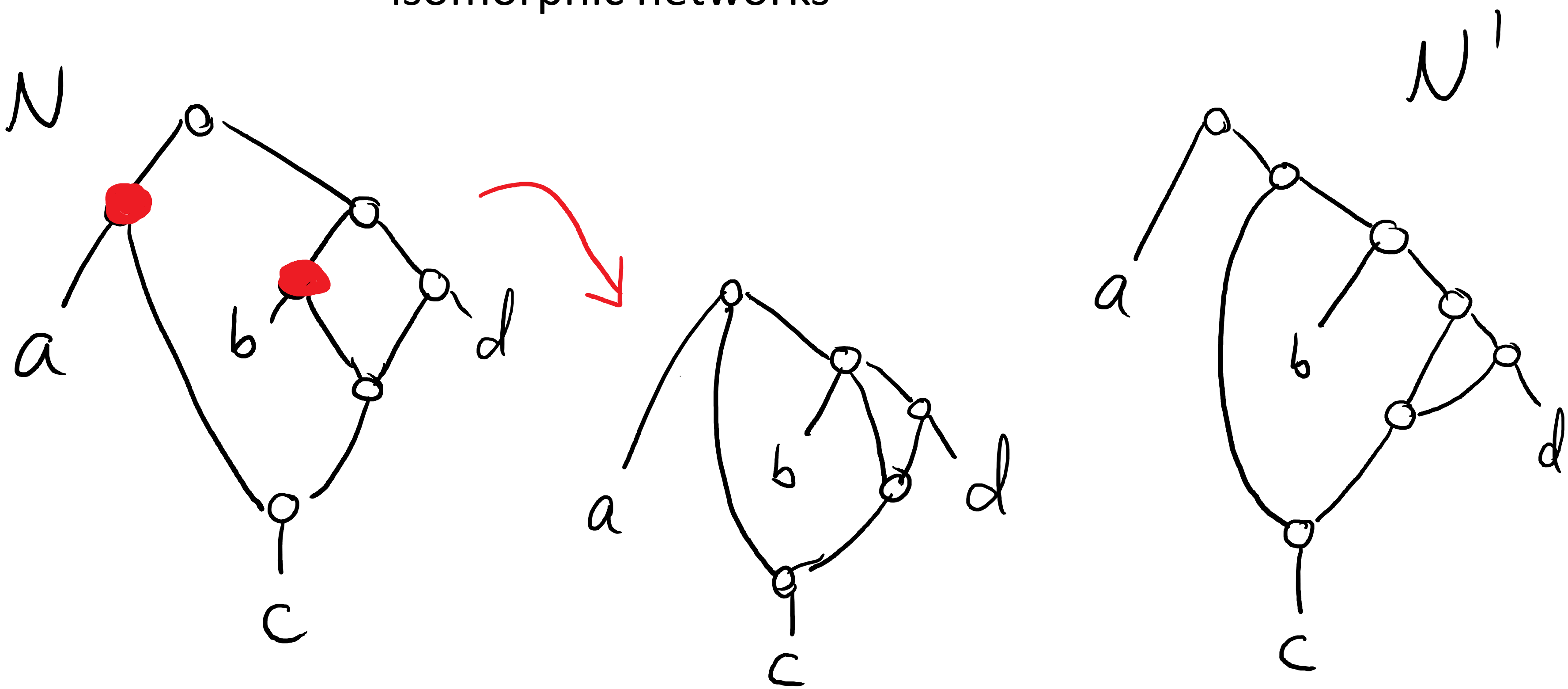




The ominus $\ominus$ operator

- 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

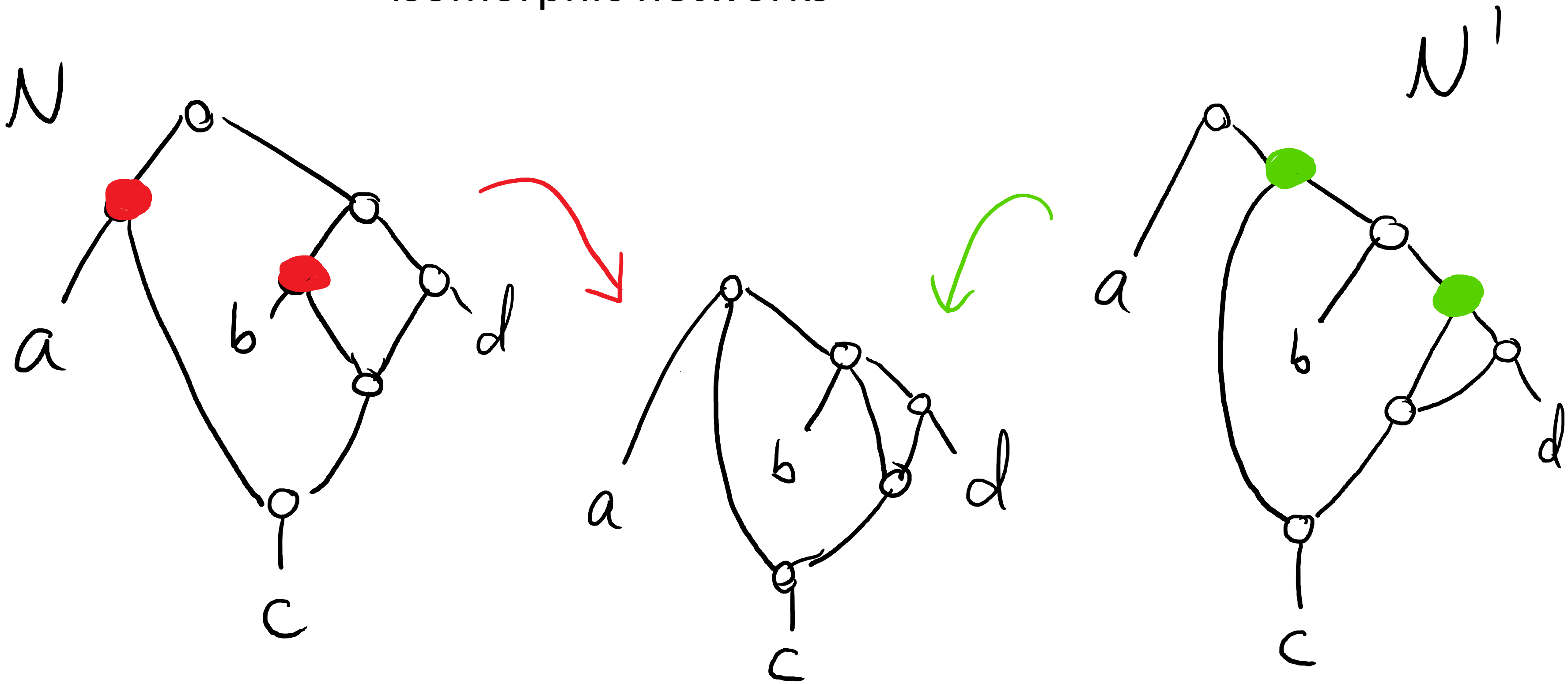
$dist_{\ominus}(N, N') =$ minimum # of $\ominus$ to apply to N and N' to obtain isomorphic networks



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

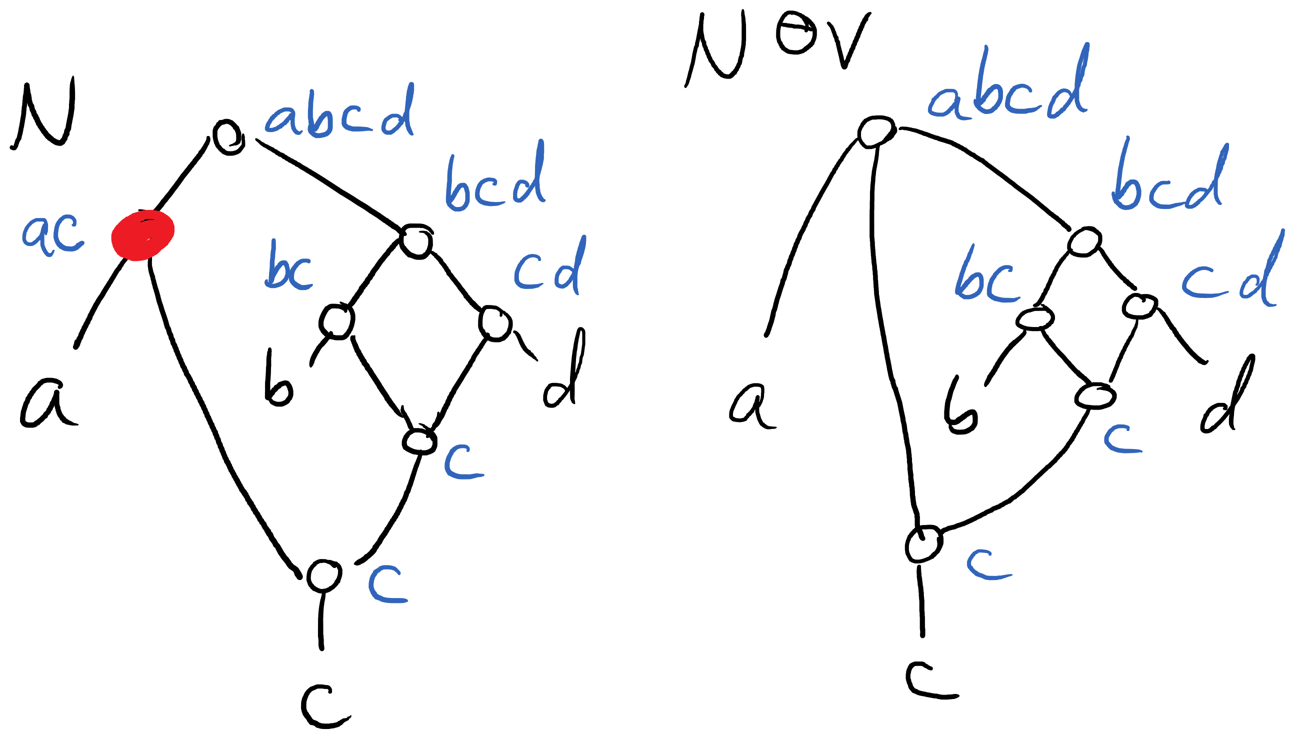
$dist_{\ominus}(N, N') =$ minimum # of $\ominus$ to apply to N and N' to obtain isomorphic networks



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

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

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



Property 2: $dist_{\ominus}(N, N') \geq mRF(N, N') \geq RF(N, N')$

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

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

Property 2: $dist_{\ominus}(N,N') \geq mRF(N,N') \geq RF(N,N')$

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

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

Both RF and mRF can be 0 on distinct networks.

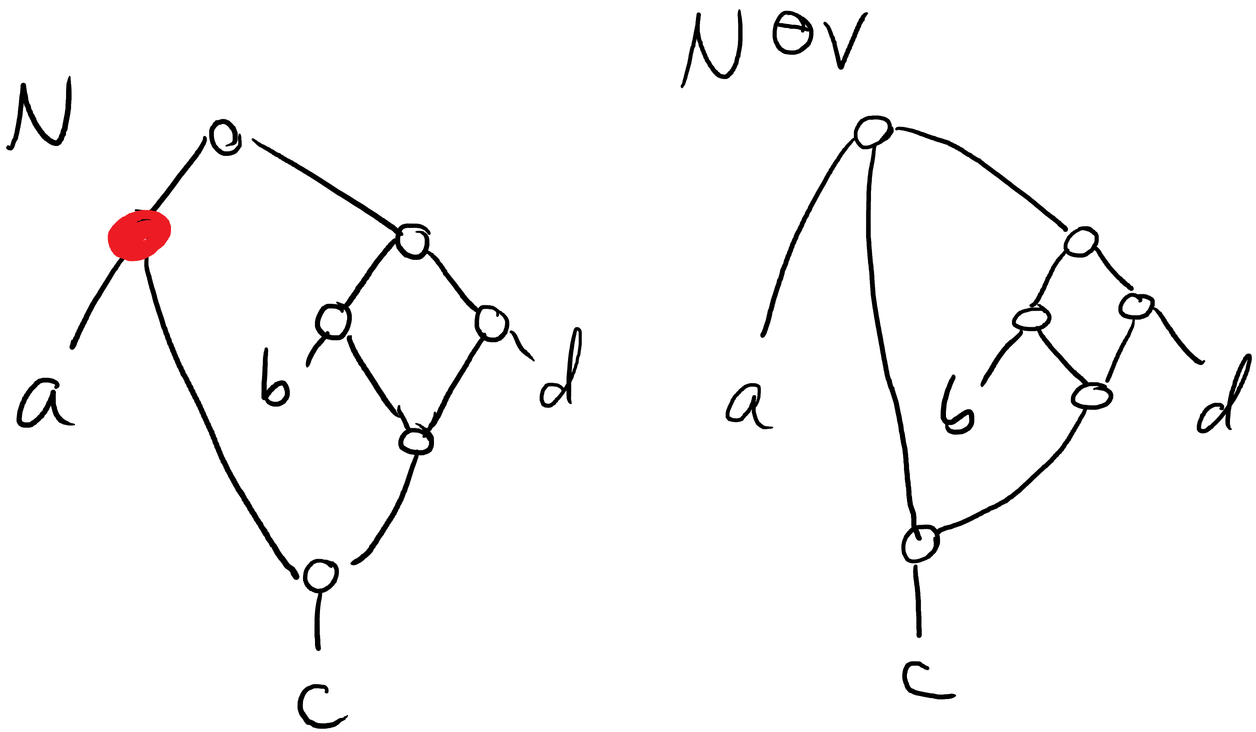
Not $dist_{\ominus}$:)

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"

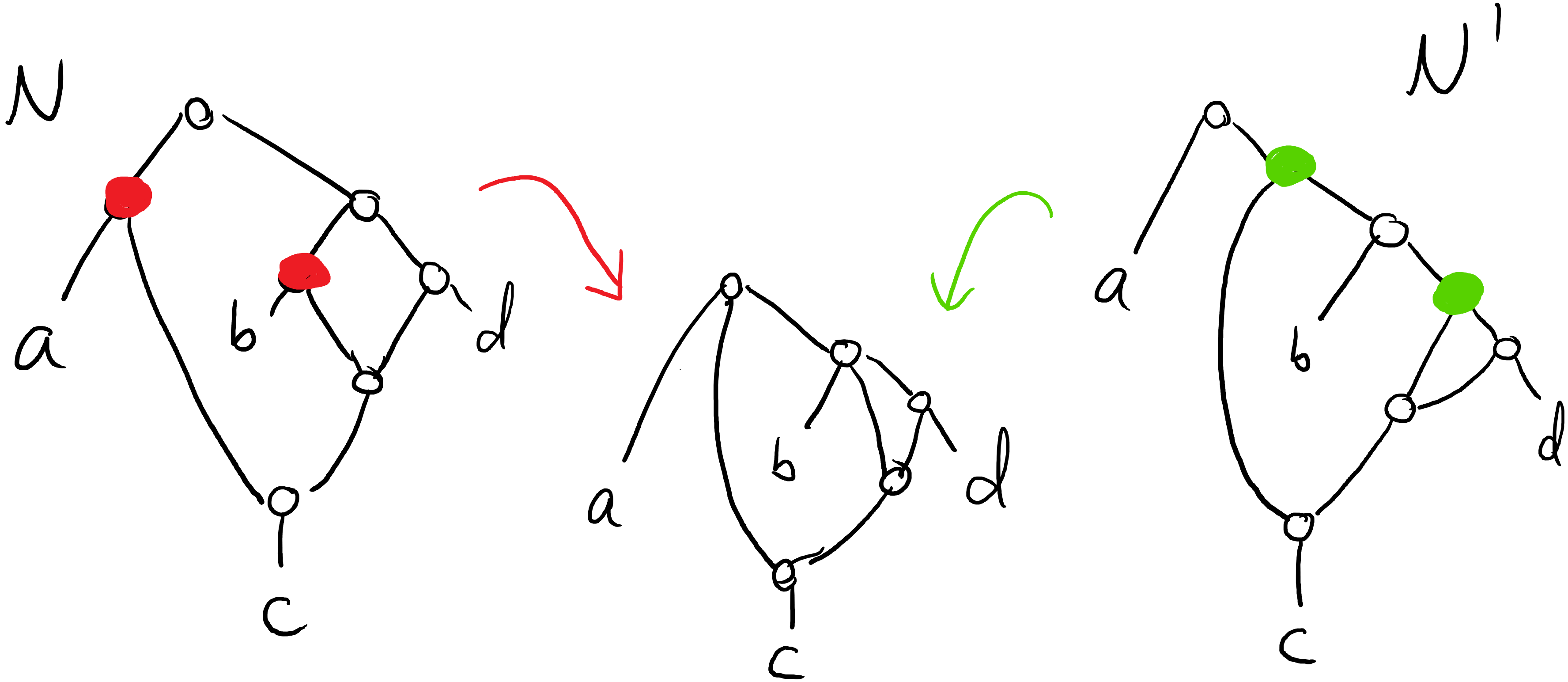


Property 3: $\ominus$ preserves ancestry relations

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.

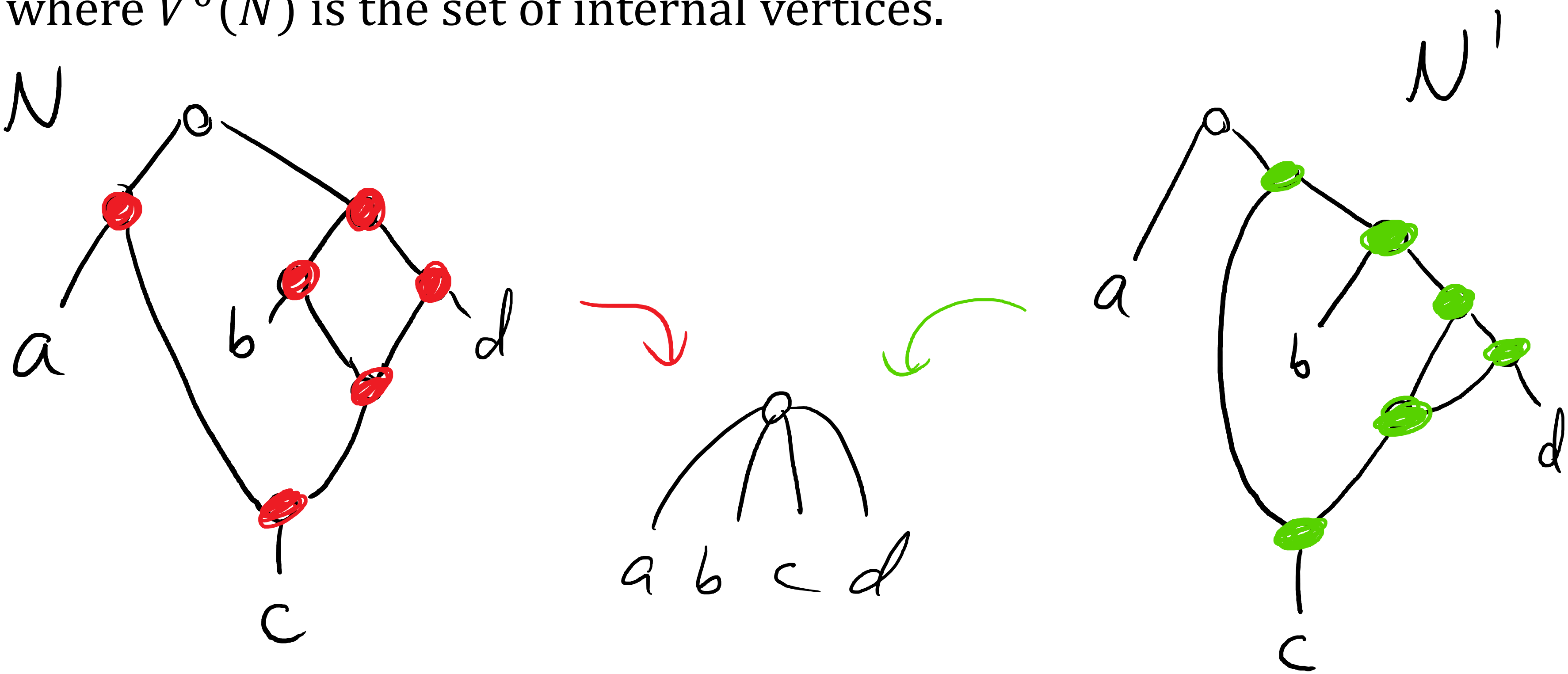
Property 5: $dist_{\ominus}(N, N')$ is a **metric** on the space of **all networks** with the same leaves.

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.

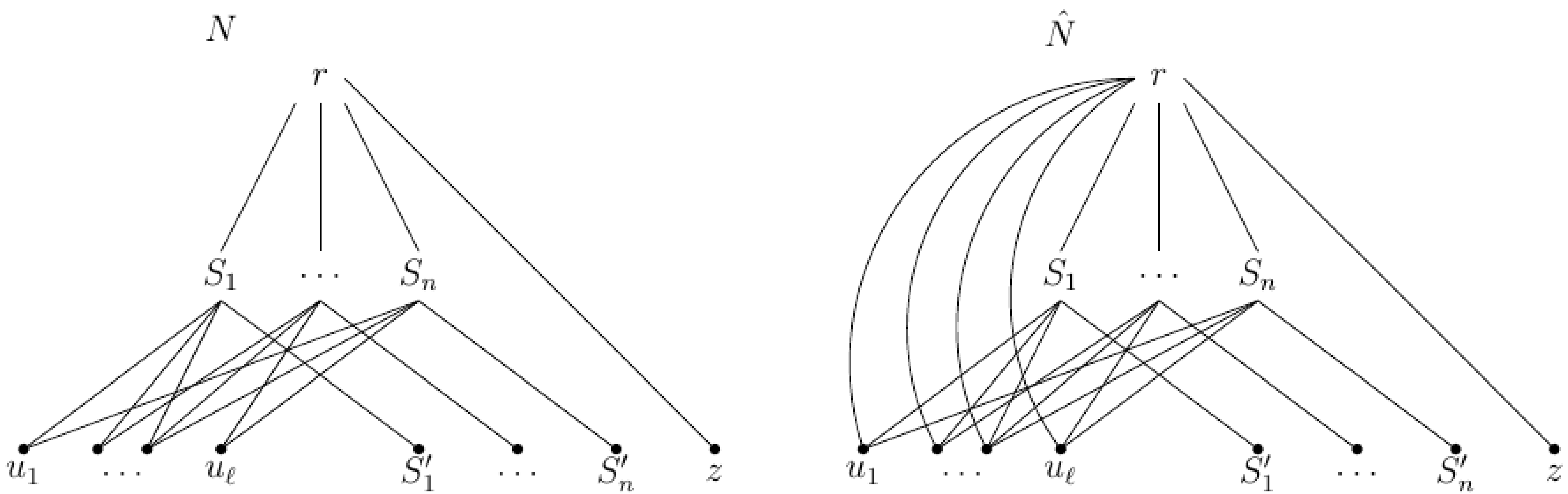
Distinct-cluster means that each cluster has multiplicity 1.

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.

SET-COVER: choose min # of sets from $S_1, \ldots, S_n$ to cover the universe.

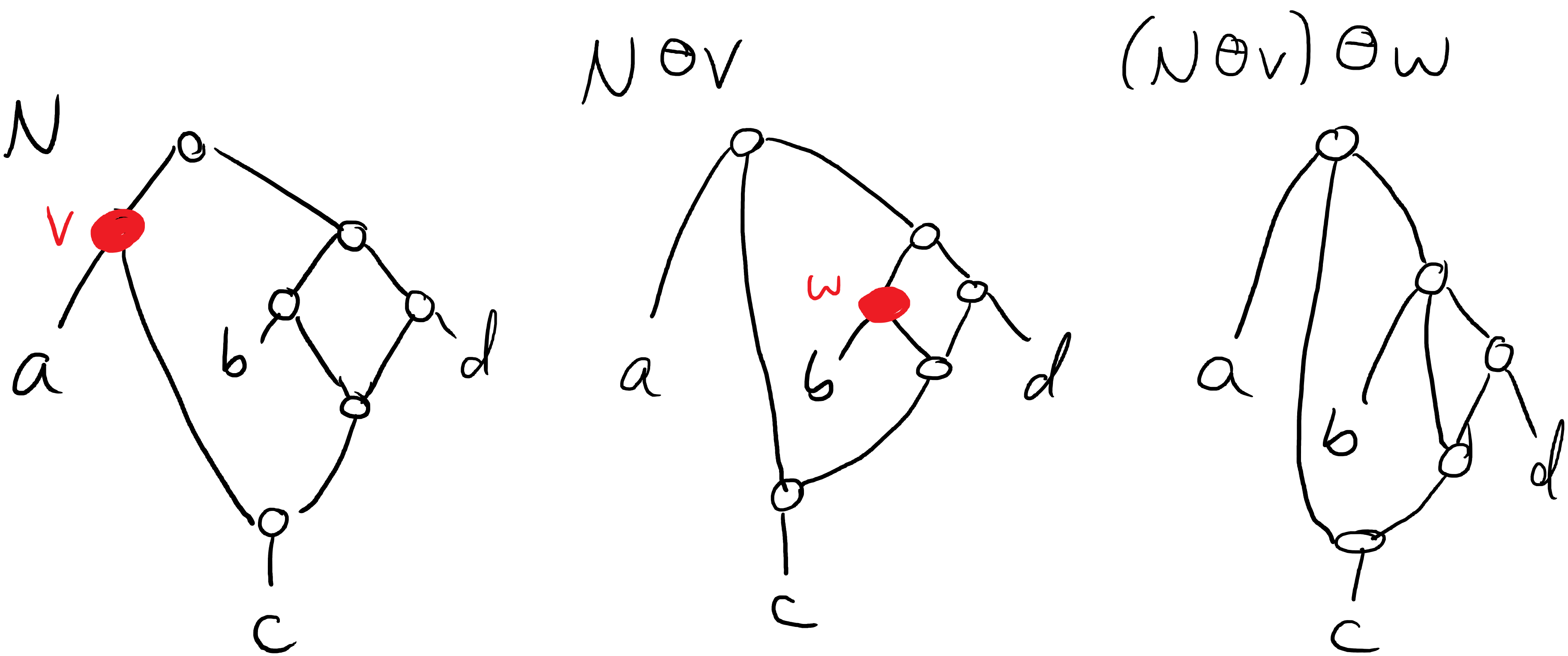
The reduction to computing $dist_{\ominus}$.



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.

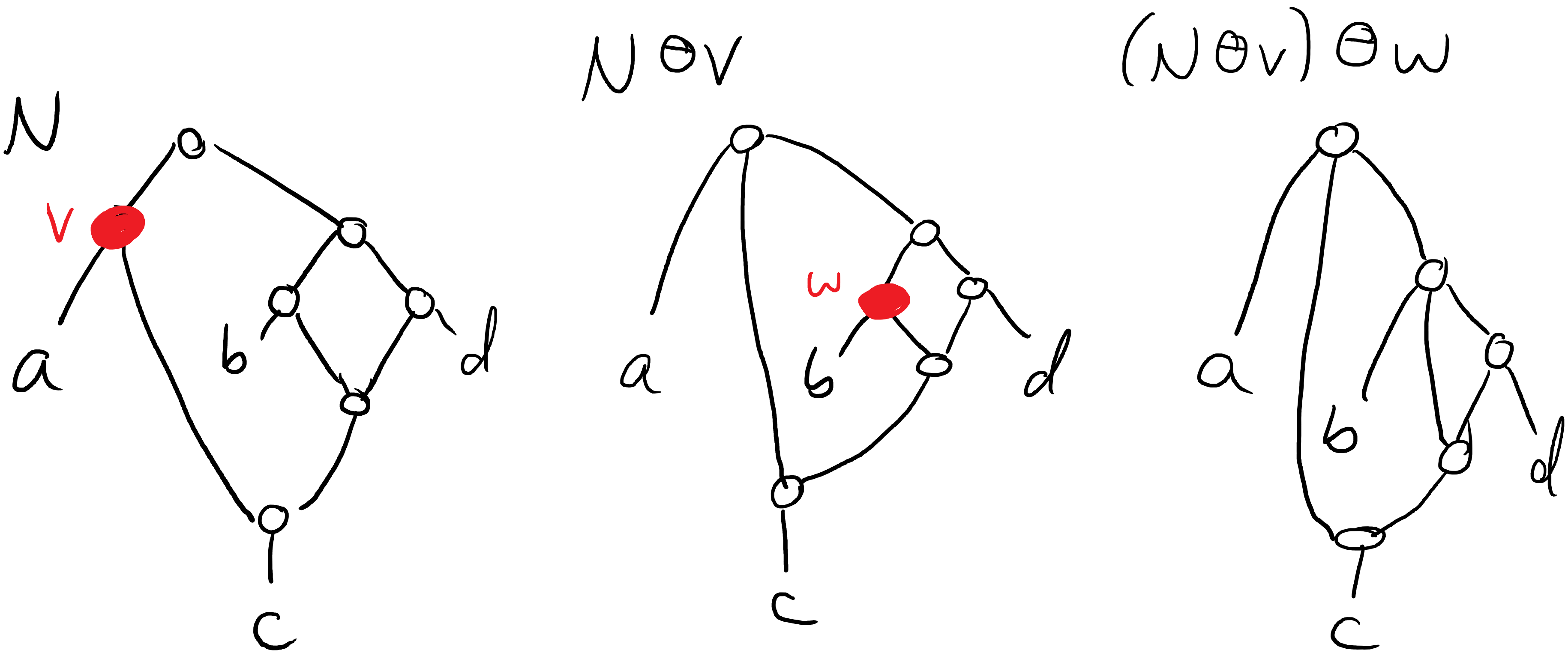


Property 6: $\ominus$ can create new shortcuts

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

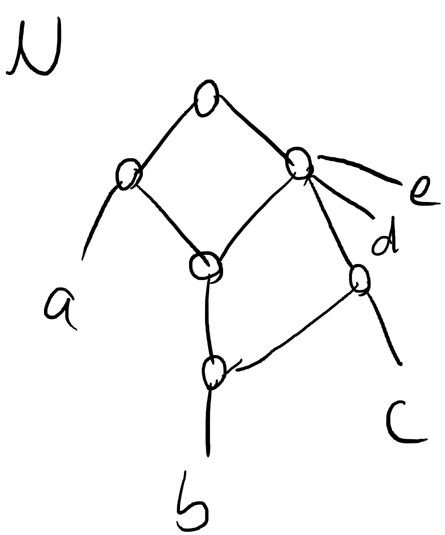
$$dist^{-}_{\ominus}(N, N') = min\{|W| + |W'| : N \ominus W \simeq_{SF} N' \ominus W'\}$$

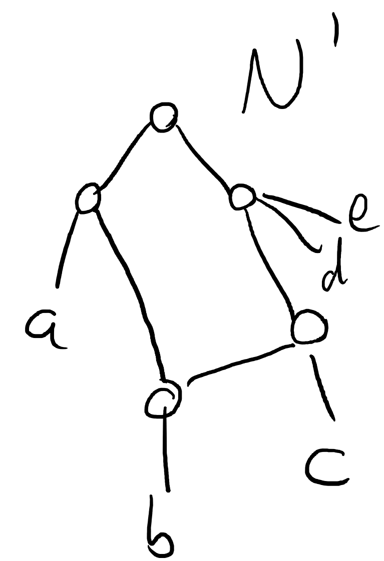
where $\simeq_{SF}$ means "isomorphic after removing shortcuts".

The shortcut-free ⊖-distance

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

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

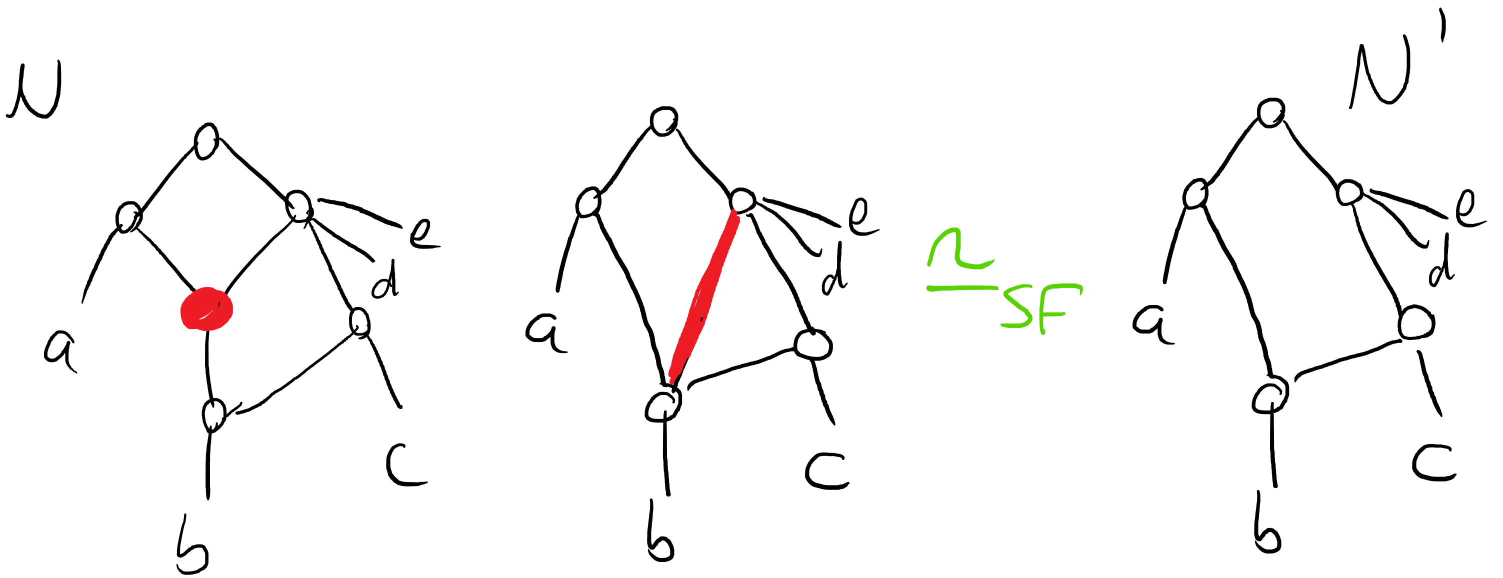




The shortcut-free $\ominus$-distance

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

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



The shortcut-free $\ominus$-distance

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

All previous properties maintained.

- $dist_{\ominus}^{-}(N, N') \geq mRF(N, N') \geq RF(N, N')$
- $dist_{\ominus}^{-}(N, N')$ 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. $dist_{\ominus}^{-}(N, N) \leq d(N, N'') + d(N'', N')$

$$dist_{\ominus}^{-}(N, N') \geq mRF(N, N')$$

$dist^{-}_{\ominus}(N, N') \geq mRF(N, N')$

When does equality hold? i.e.,

When is $dist^{-}_{\ominus}(N, N')$ just the size of the symmetric difference between the multisets of clusters of N and N'?

$dist_{\ominus}^{-}(N, N') \geq mRF(N, N')$

When does equality hold? i.e.,

When is $dist_{\ominus}^{-}(N, N')$ just the size of the symmetric difference between the multisets of clusters of N and N'?

Proposition: let Γ be a class of networks. If:

- Γ is closed under $\ominus$ and removing shortcuts, and
- every network of Γ minus shortcuts is encoded by its multisets of clusters,

then $dist_{\ominus}^{-}(N, N') = mRF(N, N')$.

Proposition: let Γ be a class of networks. If:

- Γ is closed under $\ominus$ and removing shortcuts, and
- every network of Γ minus shortcuts is encoded by its multisets of clusters,

then $dist_{\ominus}^{-}(N, N') = mRF(N, N')$.

Proposition: let Γ be a class of networks. If:

- Γ is closed under $\ominus$ and removing shortcuts, and
- every network of Γ minus shortcuts is encoded by its multisets of clusters,

then $dist_{\ominus}^{-}(N, N') = mRF(N, N')$.

Theorem: $dist_{\ominus}^{-}(N, N') = mRF(N, N')$ 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: for N, N' distinct-cluster networks as above, 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.

Definition: let N, N' be two **distinct-cluster networks** with $V(N) = V(N')$, and such that each vertex has the same cluster in either network.

A **bypassing set** is a set of vertices $W \subseteq V(N)$ such that for any $u, v \in V(N) \setminus W$, we have $u \preccurlyeq_N v \Leftrightarrow u \preccurlyeq_{N'} v$.

- In other words, a bypassing set intersect all conflicting ancestry relations.

Definition: let N, N' be two **distinct-cluster networks** with $V(N) = V(N')$, and such that each vertex has the same cluster in either network.

A **bypassing set** is a set of vertices $W \subseteq V(N)$ such that for any $u, v \in V(N) \setminus W$, we have $u \preccurlyeq_N v \Leftrightarrow u \preccurlyeq_{N'} v$.

- In other words, a bypassing set intersect all conflicting ancestry relations.

Theorem: for N, N' distinct-cluster networks as above, $dist_{\ominus}^{-}(N, N')$ is twice the minimum size of a bypassing set.

Theorem: for N, N' distinct-cluster networks as above, $dist^{-}_{\ominus}(N, N')$ is twice the minimum size of a bypassing set.

Bypassing sets need to destroy at least one vertex per bad ancestry relation.

This is just VERTEX-COVER on the graph of bad ancestry relations.

Corollary: for N, N' distinct-cluster networks as above, $dist^{-}_{\ominus}(N, N')$ 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.

Theorem: for N, N' distinct-cluster networks as above, $dist_{\ominus}^{-}(N, N')$ is twice the minimum size of a bypassing set.

Bypassing sets need to destroy at least one vertex per bad ancestry relation.

This is just VERTEX-COVER on the graph of bad ancestry relations.

Corollary: for N, N' distinct-cluster networks as above, $dist_{\ominus}^{-}(N, N')$ 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.

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

Corollary: for N, N' distinct-cluster networks as above, $dist_{\ominus}^{-}(N, N')$ 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.

Behaves very differently than $dist_{\ominus}$.

However,

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

Corollary: for N, N' distinct-cluster networks as above, $dist^{-}_{\ominus}(N, N')$ 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.

Behaves very differently than $dist_{\ominus}$

However,

Theorem: f ... mputing $dist^{-}_{\ominus}(N$

I was fully expecting to skip these slides, btw

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 FPT if each cluster has multiplicity at most c?
- Any class for which $dist_{\ominus}$ is tractable?
- Characterize I called bad ancestry graphs?

- Implement and evaluate computation of $dist_{\ominus}$ and $dist_{\ominus}^{-}$