New reduction rules (and tight kernels) for computing the distance between phylogenetic trees

Steven Kelk

Department of Data Science and Knowledge Engineering
Maastricht University
The Netherlands

Email: steven.kelk@maastrichtuniversity.nl

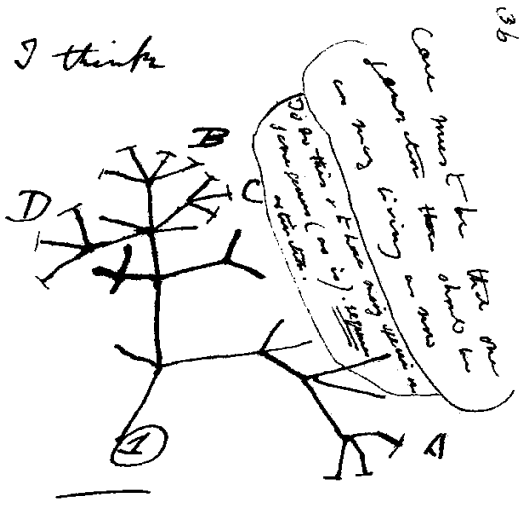
Joint work with Simone Linz (Auckland)

Oxford University, 13th June 2019.

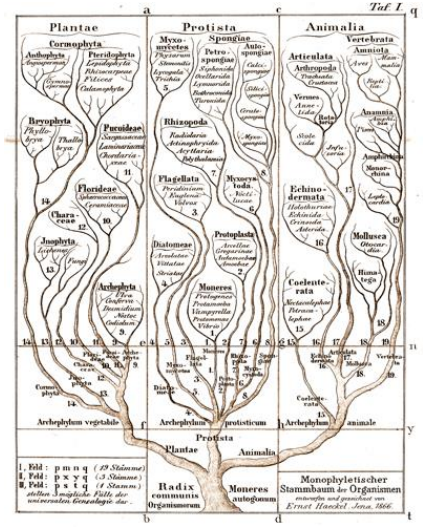
Maastricht University

Phylogenetics

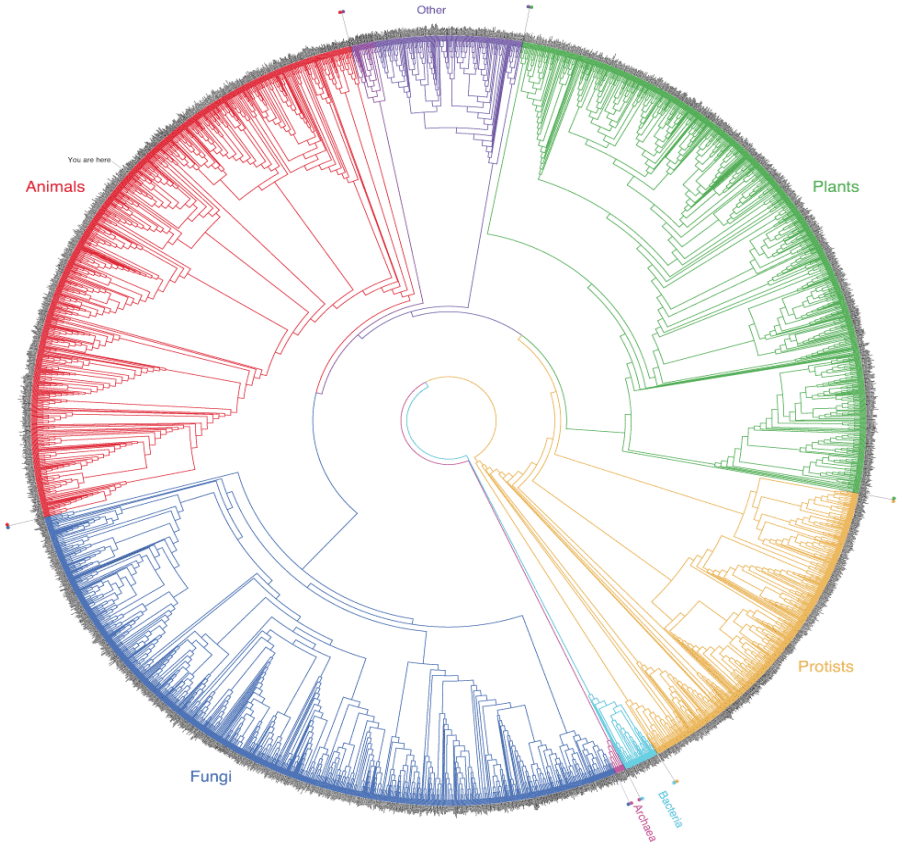
The reconstruction and analysis of evolutionary trees and networks based on molecular sequence data or morphological characters.



Charles Darwin (1837)

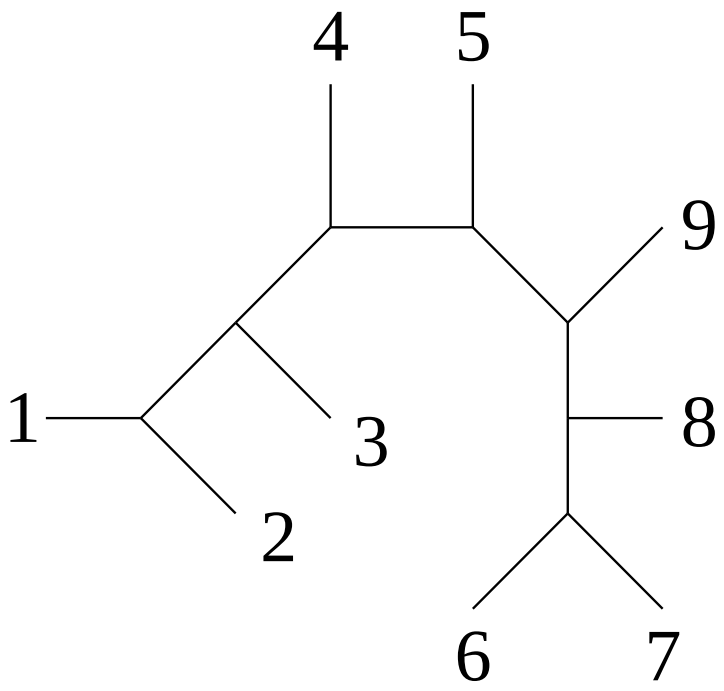


Ernst Haeckel (1866)



Hillis, et al. (2003)

Phylogenetic trees



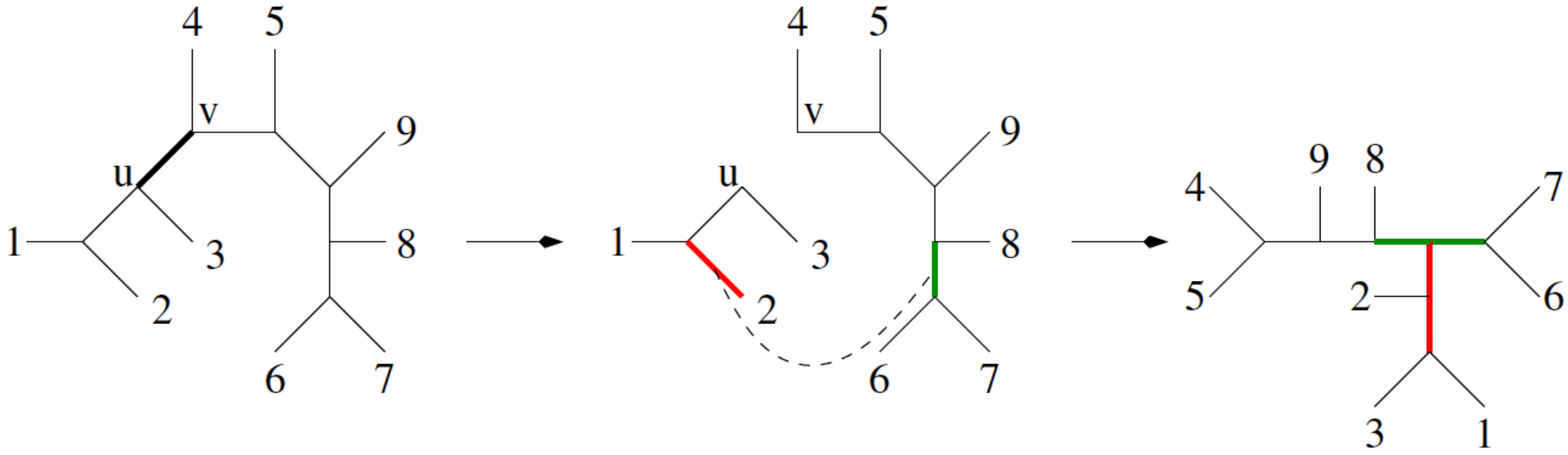
An (*unrooted) phylogenetic tree on X* is a connected acyclic graph whose internal vertices have degree three and whose leaf set is *X*.

Distances between phylogenetic trees

We wish to compare two trees, i.e. to quantify the dissimilarities between them.

Distances between trees provide a lower bound on the number of non-tree-like events, such as hybridization, which can cause the topologies of the trees to differ.

Tree bisection and reconnection (TBR)



Let $d_{TBR}(T,T')$ denote the minimum number of TBR operations required to transform T into T'. Then, $d_{TBR}(T,T')$ induces a metric on the space of all unrooted phylogenetic trees with n leaves.

(Robinson, 1971; Allen and Steel, 2001).

Computing $d_{TBR}(T,T')$ is NP-hard and fixed-parameter tractable, when parameterized by $k=d_{TBR}$.

(Hein et al., 1996; Allen and Steel, 2001).

Fixed-parameter tractability of d_{TBR}

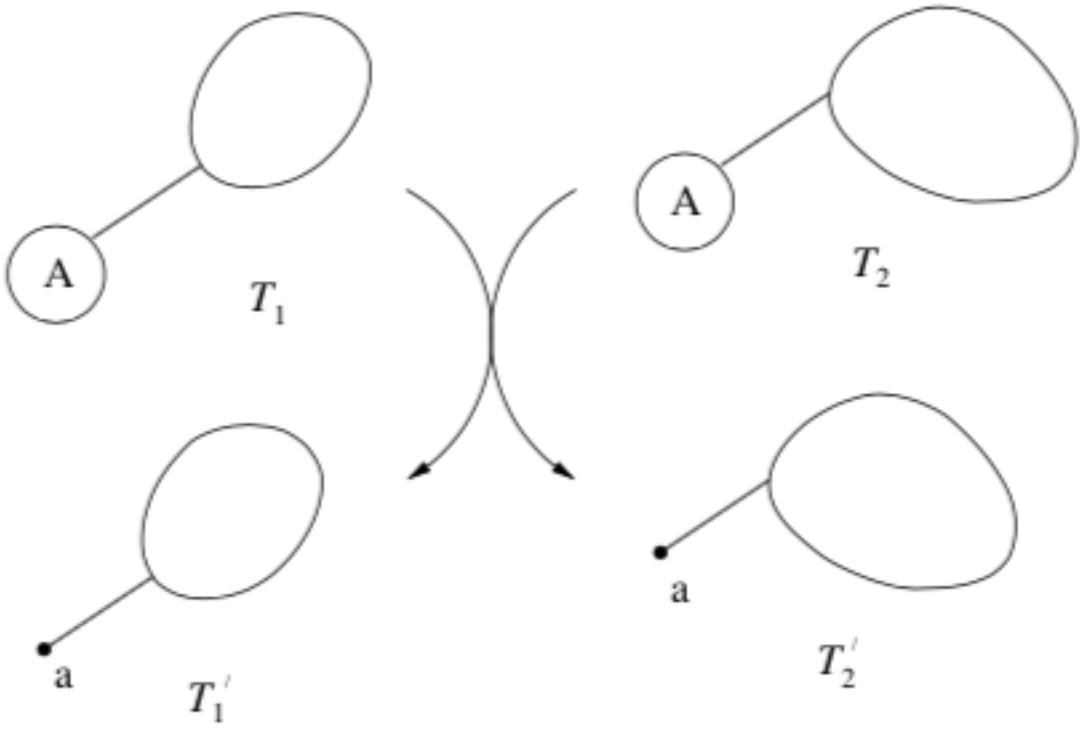
Kernelization idea (Allen and Steel 2001): Shrink two trees to their common cores by applying two reduction rules. Then show that

- the two reductions preserve the TBR distance, and
- the size (i.e. number of leaves) of the two smaller trees is bounded from above by a function that is linear in the TBR distance.

Time to decide if $d_{TBR}(T,T')$ is at most k is

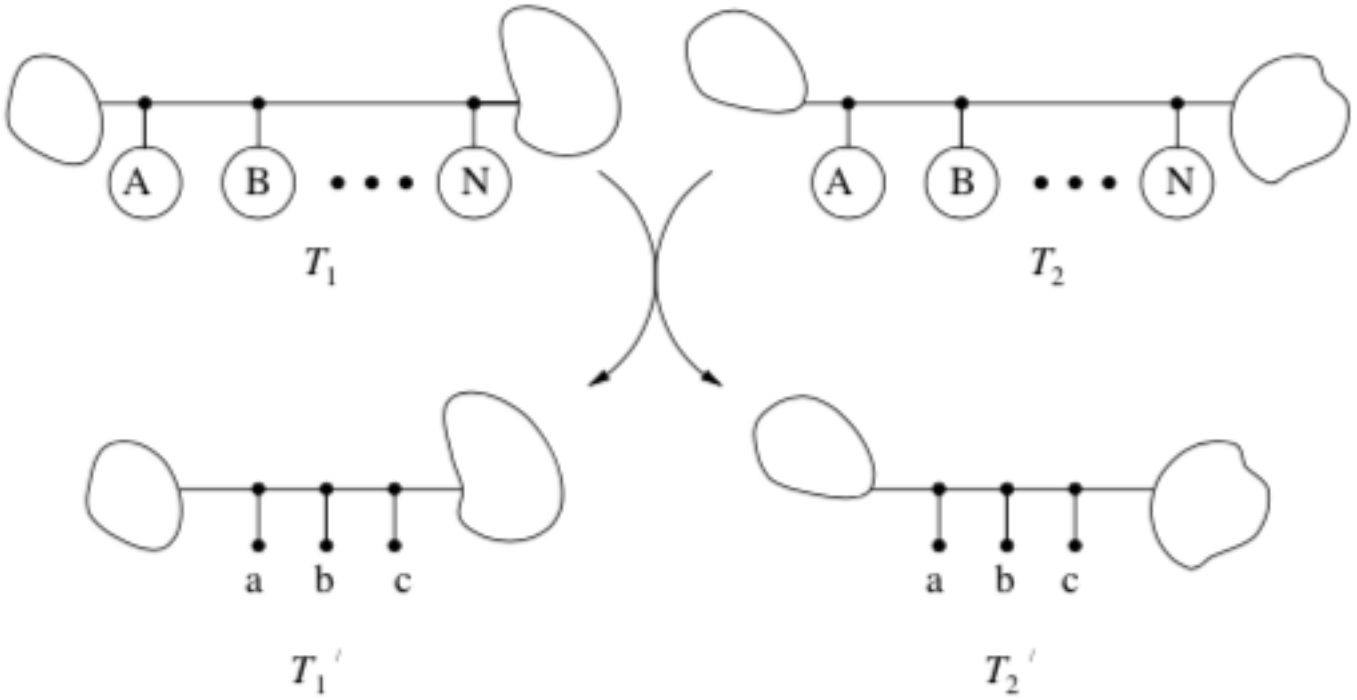
$$O(f(k) + p(n)).$$

Subtree reduction



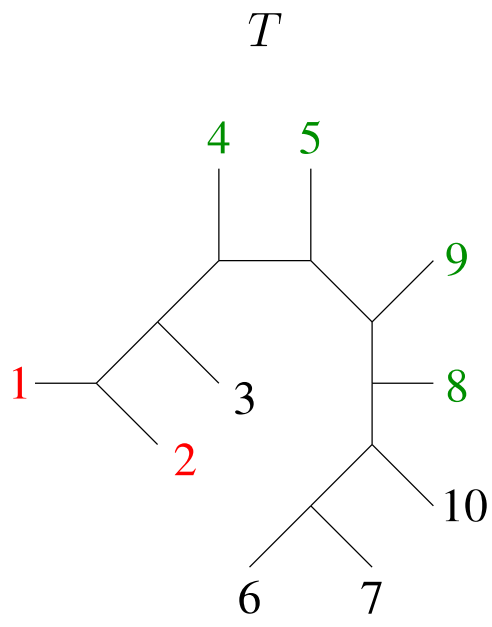
Allen and Steel, 2001

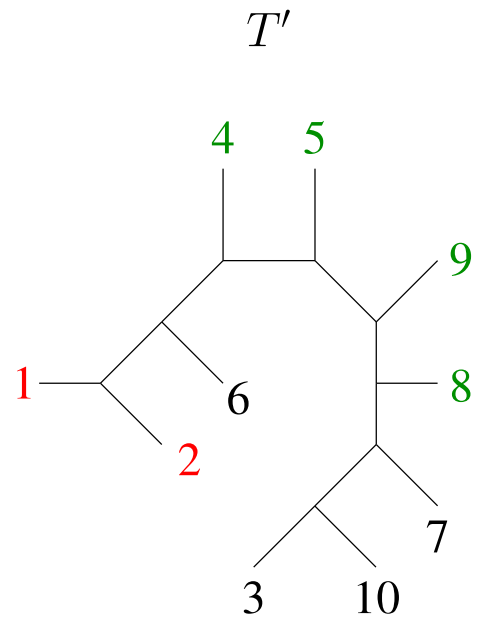
Chain reduction

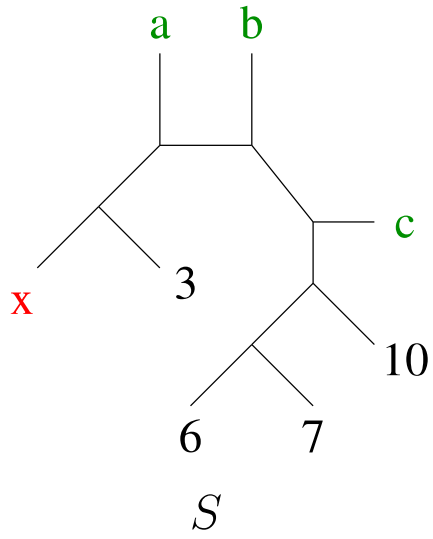


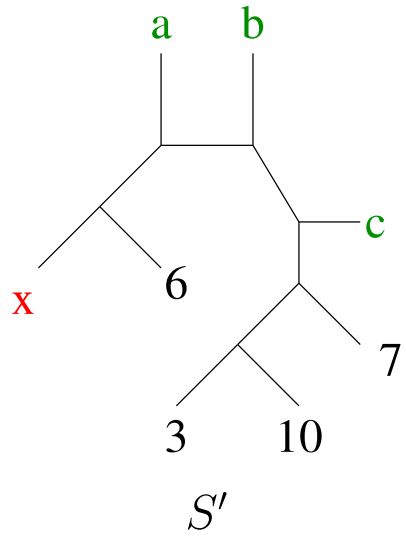
Allen and Steel, 2001

Example









Theorem. (Allen and Steel, 2001).

[Reductions are safe] Let S and S' be two trees obtained from T and T' by applying a single subtree or chain reduction. Then

$$d_{\mathrm{TBR}}(T,T') = d_{\mathrm{TBR}}(S,S').$$

[Linear kernel] Let S and S' be two trees obtained from T and T' by repeated applications of the subtree and chain reduction until no further reduction is possible. Then

$$|X'| \leq 28 d_{\mathrm{TBR}}(T,T'),$$

where X' is the leaf set of S and S'.

Theorem. (Allen and Steel, 2001).

[Reductions are safe] Let S and S' be two trees obtained from T and T' by applying a single subtree or chain reduction. Then

$$d_{\mathrm{TBR}}(T,T') = d_{\mathrm{TBR}}(S,S').$$

[Linear kernel] Let S and S' be two trees obtained from T and T' by repeated applications of the subtree and chain reduction until no further reduction is possible. Then

$$|X'| \leq 28 d_{\mathrm{TBR}}(T,T'),$$

where X' is the leaf set of S and S'.

How good is this bound/is it tight?

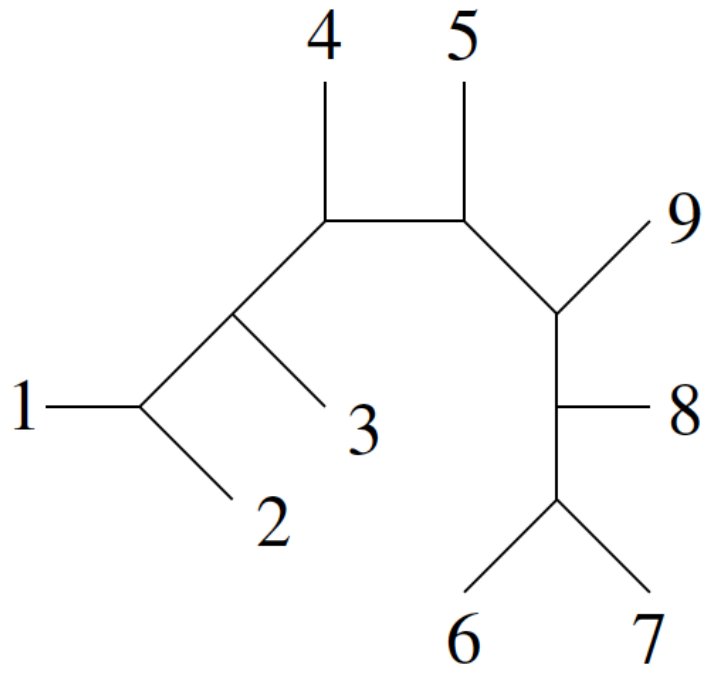
Today we show two things:

- We reanalyse Allen and Steel's kernel, and show that it is considerably smaller than they claimed: $15d_{\text{TBR}} - 9$. Moreover, this is tight.

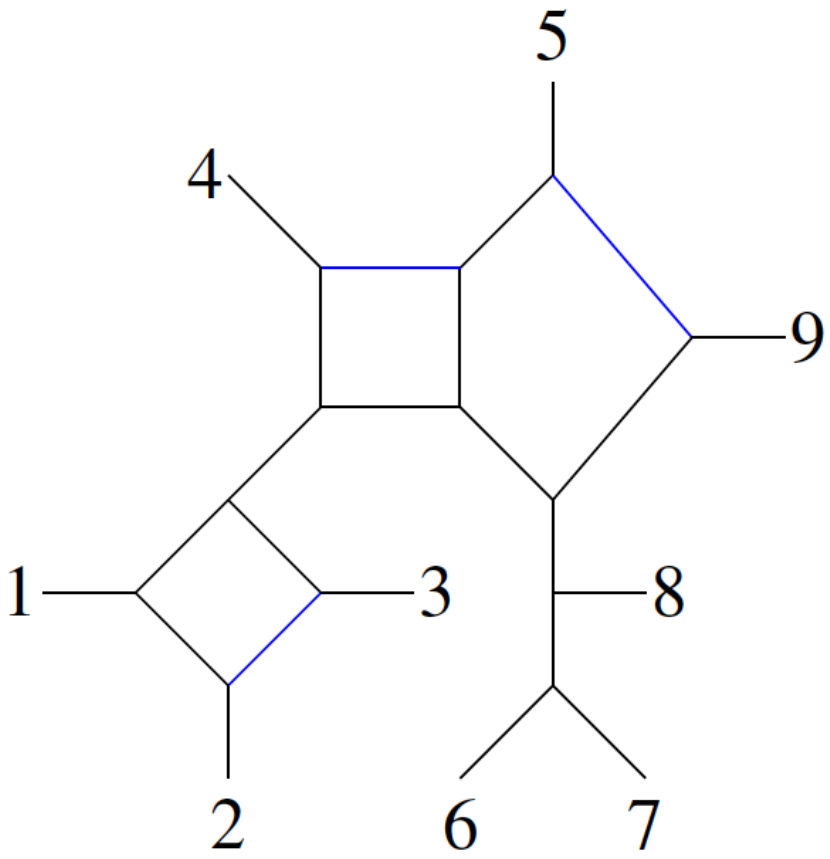
- We devise a number of new reduction rules which, when combined with Allen and Steel's reduction rules, yield a kernel of size: $11d_{\text{TBR}} - 9$. This is also tight.

Strategy to achieve $15d_{\mathrm{TBR}} - 9$: We translate the problem of computing d_{TBR} into a problem on phylogenetic *networks* (i.e. graphs), establish a smaller kernel that is based on the same two reductions, and show that this new kernel is tight.

Phylogenetic networks to the rescue

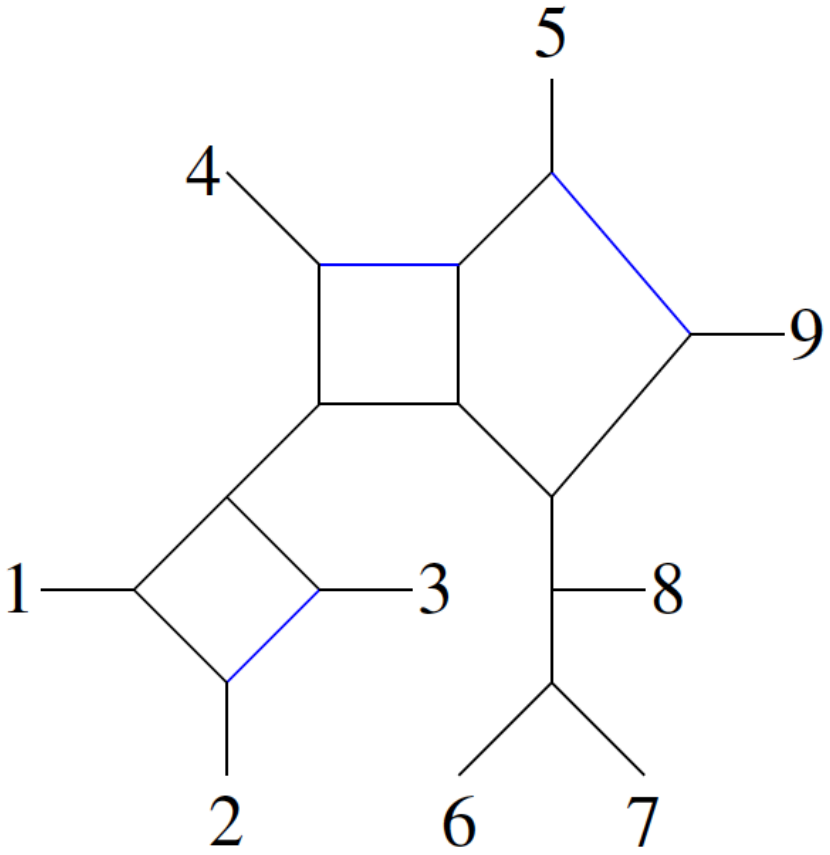


An *unrooted phylogenetic tree on X* is a connected acyclic graph whose internal vertices have degree three and whose leaf set is *X*.



An *unrooted phylogenetic network N on X* is a simple graph whose internal vertices have degree three and whose leaf set is *X*.

Phylogenetic networks to the rescue



Reticulation number of *N* is

$r(N) = |E| - (|V|-1)$.

(equal to cyclomatic number).

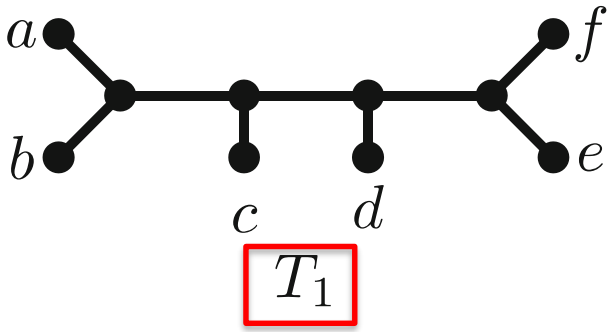
Example. $r(N) = 3$

For two trees *T* and *T'*, define the *hybridization number*

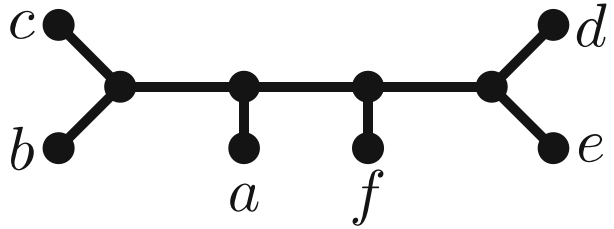
$$h^u(T, T') = \min_N \{r(N)\}$$

Where the minimum is taken over all N that embed *T* and *T'*.

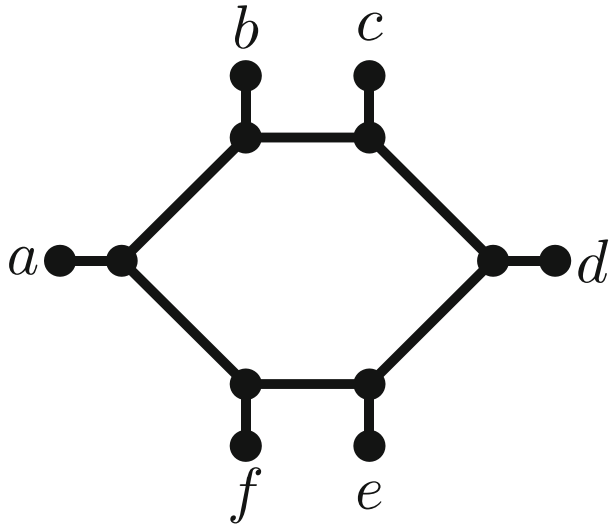
Example.



T_1



T_2



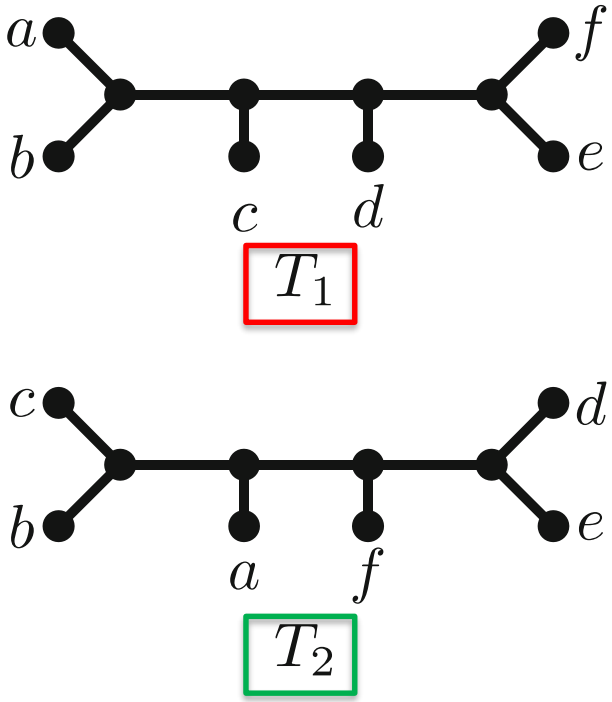
(van Iersel et al., 2018).

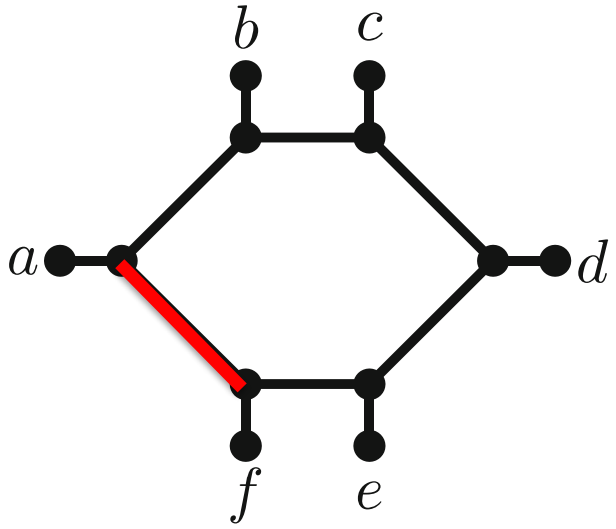
For two trees *T* and *T'*, define the *hybridization number*

$$h^u(T, T') = \min_N \{r(N)\}$$

Where the minimum is taken over all N that embed *T* and *T'*.

Example.





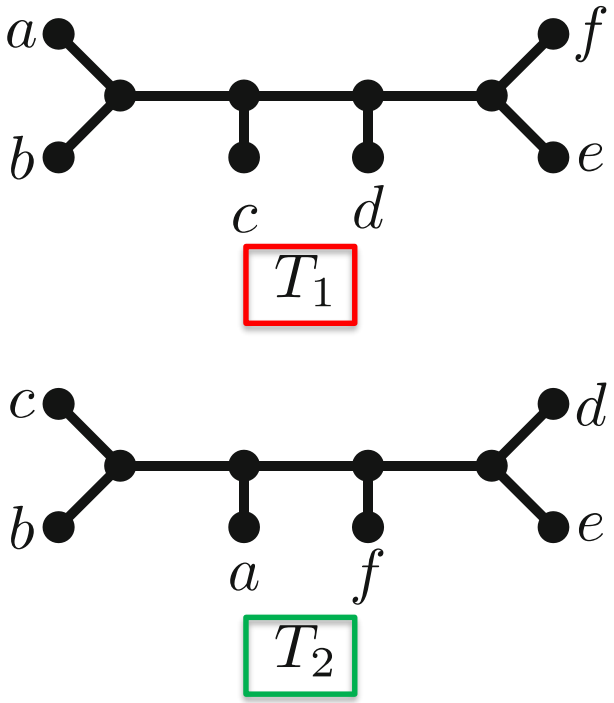
(van Iersel et al., 2018).

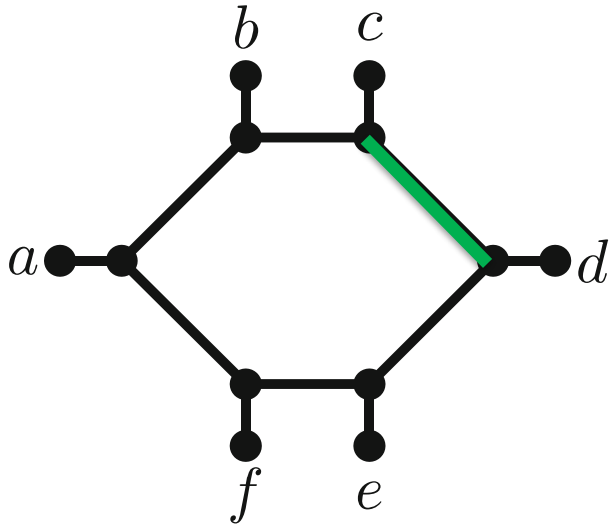
For two trees *T* and *T'*, define the *hybridization number*

$$h^u(T, T') = \min_N \{r(N)\}$$

Where the minimum is taken over all N that embed *T* and *T'*.

Example.





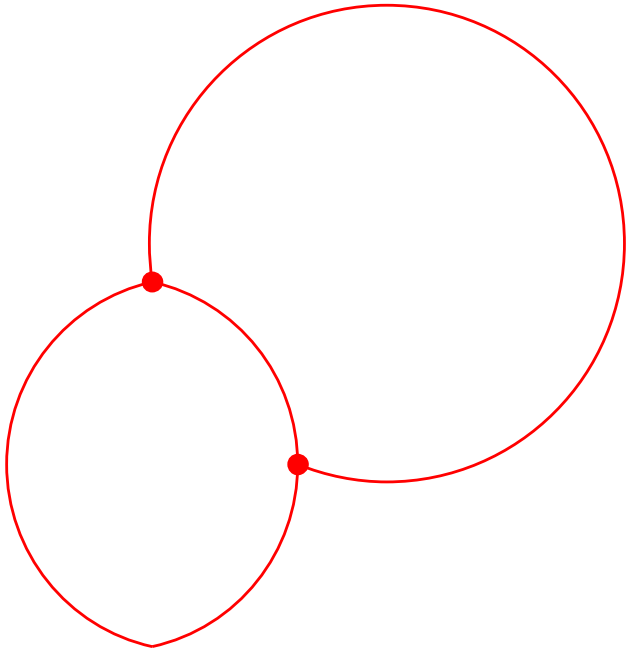
(van Iersel et al., 2018).

Theorem. (van Iersel et al., 2018)

Let T and T' be two trees. Then

$$d_{\mathrm{TBR}}(T,T') = h^{u}(T,T')$$

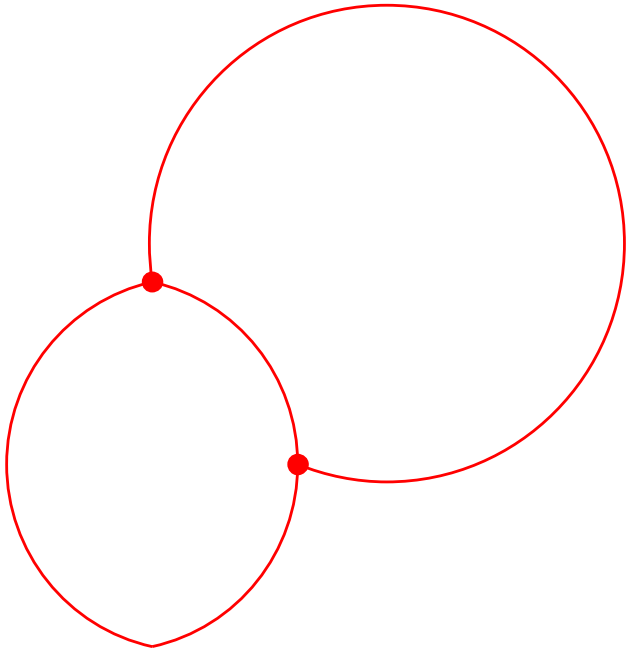
Backbones of phylogenetic networks



2-generator *G* with three edges (*sides*)

In general, for $k \geq 2$, a *k-generator* is a connected cubic multigraph such that $k = |E|-(|V|-1)$.

Backbones of phylogenetic networks

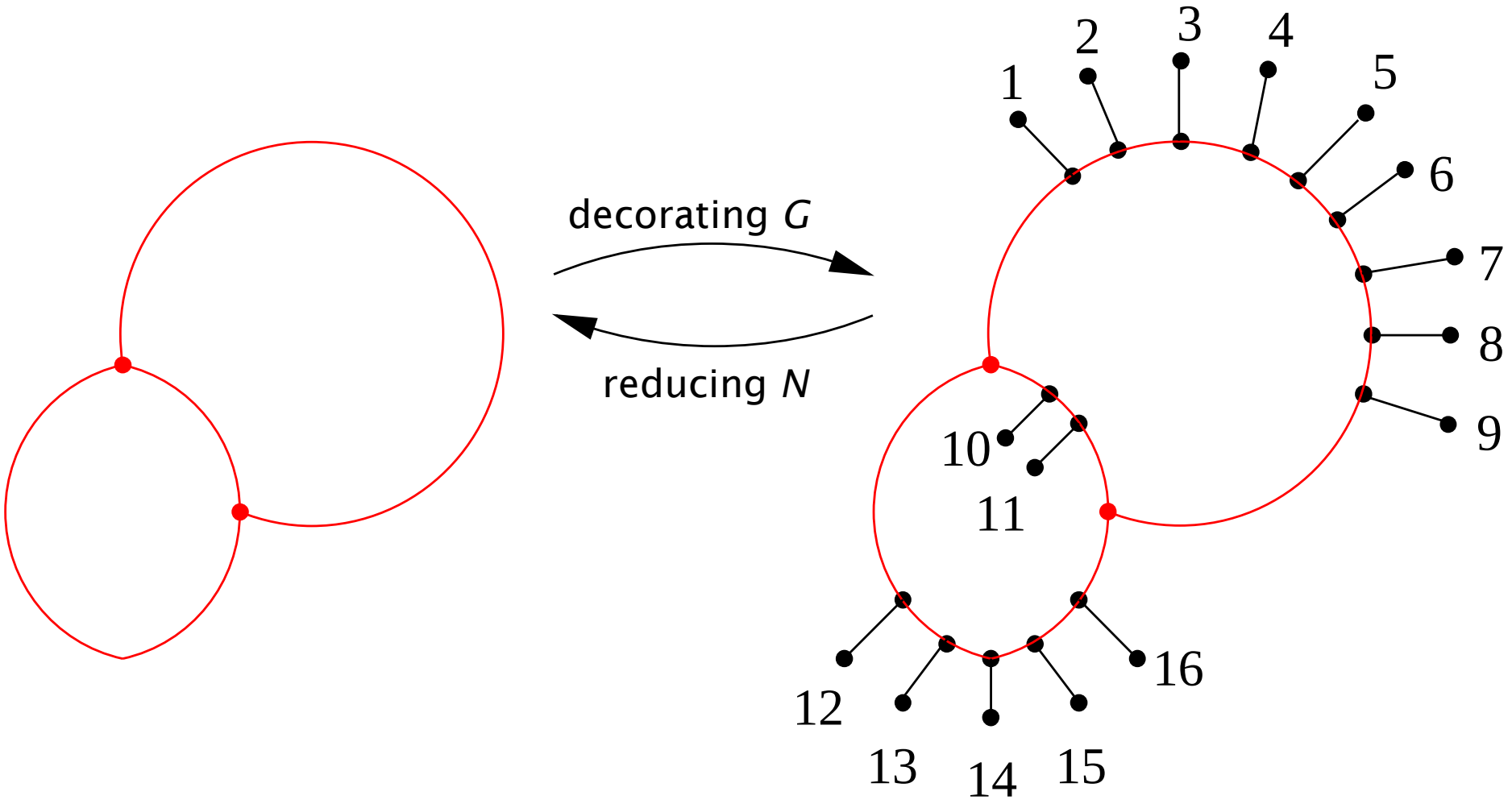


In general, for $k \geq 2$, a *k-generator* is a connected cubic multigraph such that $k = |E|-(|V|-1)$.

2-generator *G* with three edges (*sides*)

More generally: *k*-generator *G* has $3(k-1)$ edges (*sides*)

Backbones of phylogenetic networks



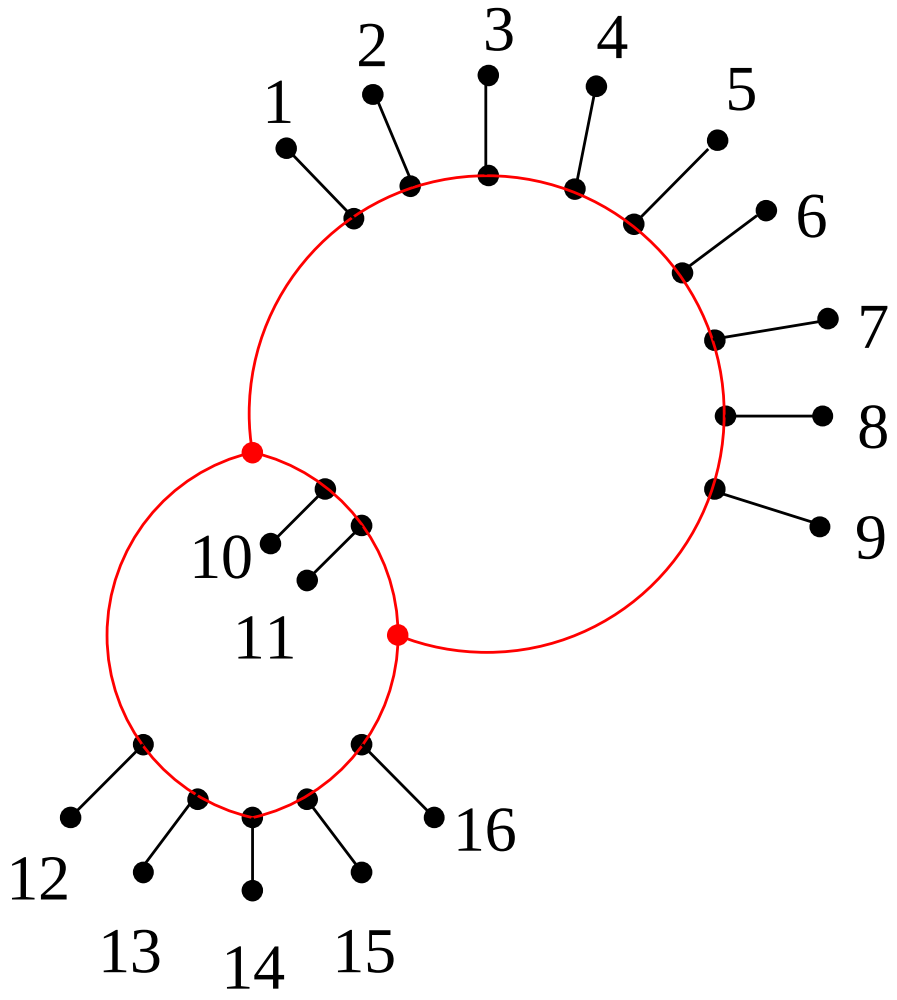
2-generator *G* with
three edges (*sides*)

network *N* with *r(N)=2*
with no pendant subtree

Backbones of phylogenetic networks

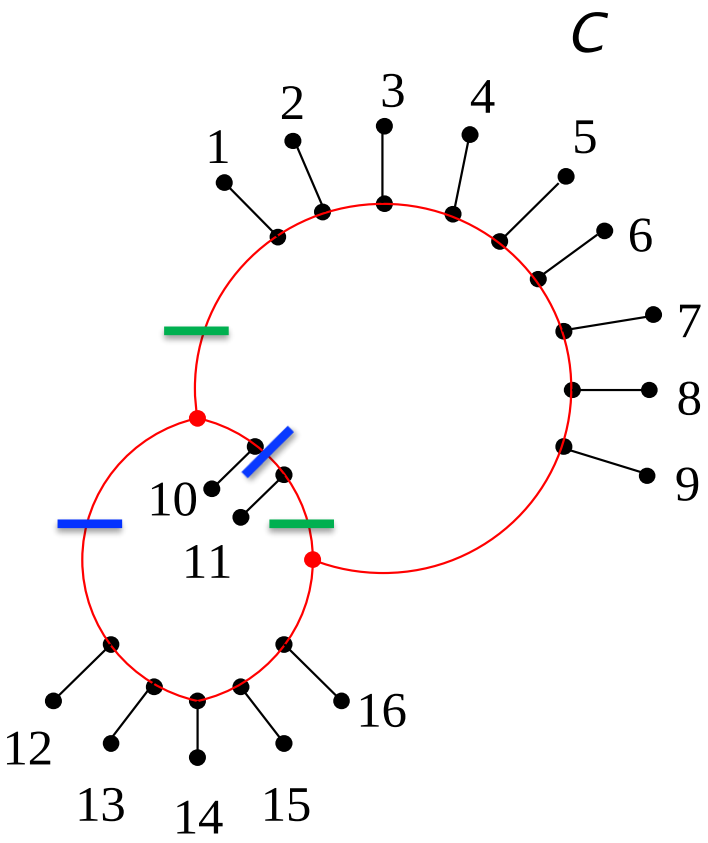
Can we bound the number of leaves that decorate a single side of *G*?

(Can we have more than 9 leaves on a side?)



network *N* with *r(N)=2* with no pendant subtree

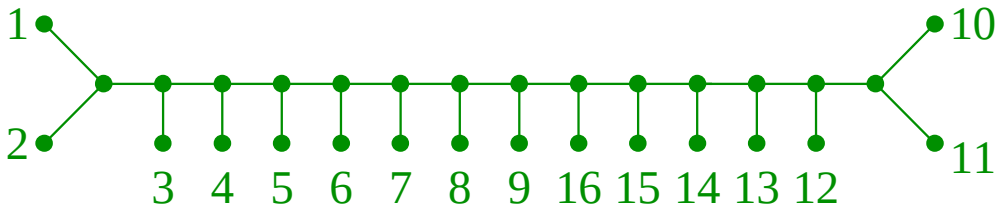
Breakpoints



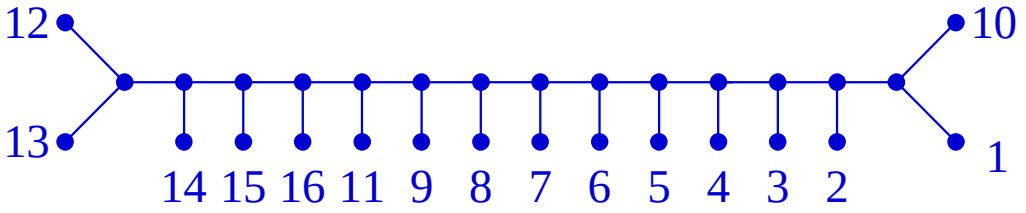
r(N)=2

Chain *C=(1,2,…,9)* has **no** breakpoint relative to *T* and *T'*.

T

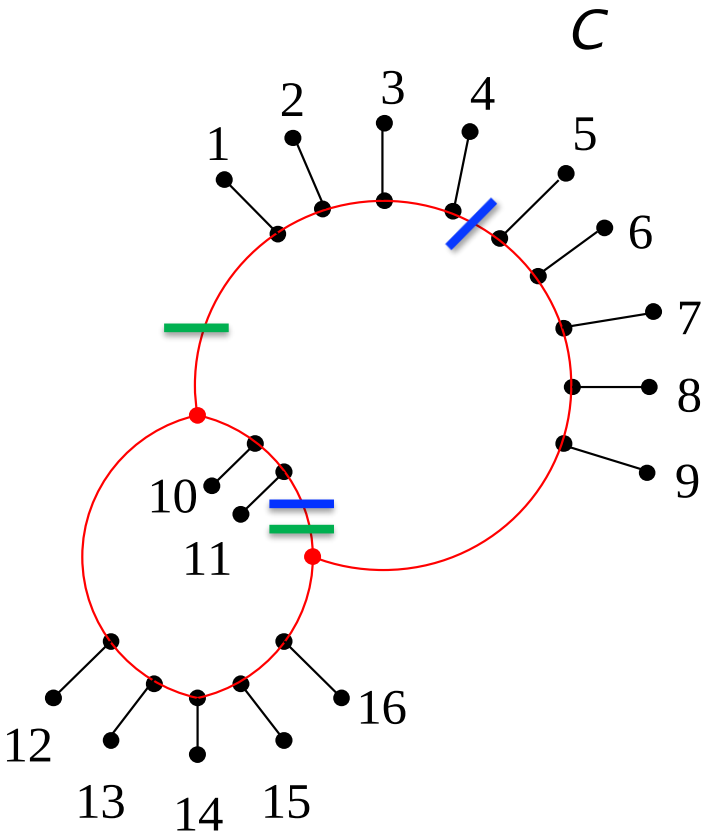


T'



C survives in *T* and *T'*.

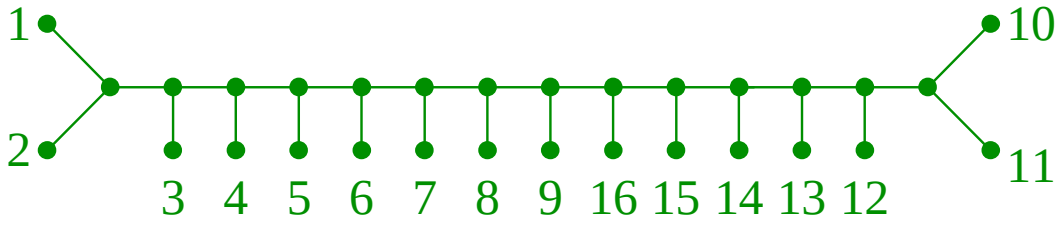
Breakpoints



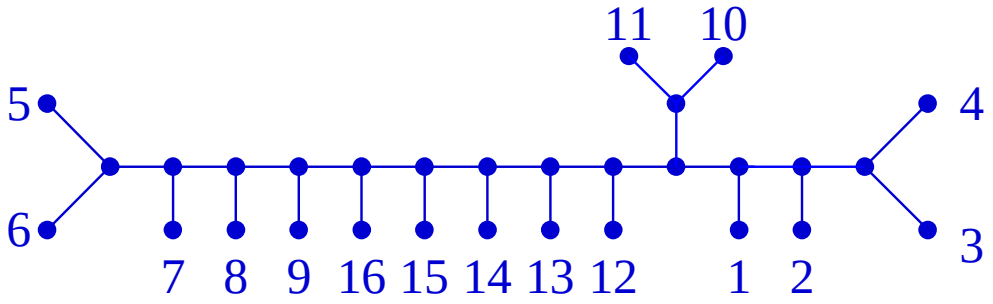
r(N)=2

Chain *C=(1,2,…,9)* has **one** breakpoint relative to *T* and *T'*.

T

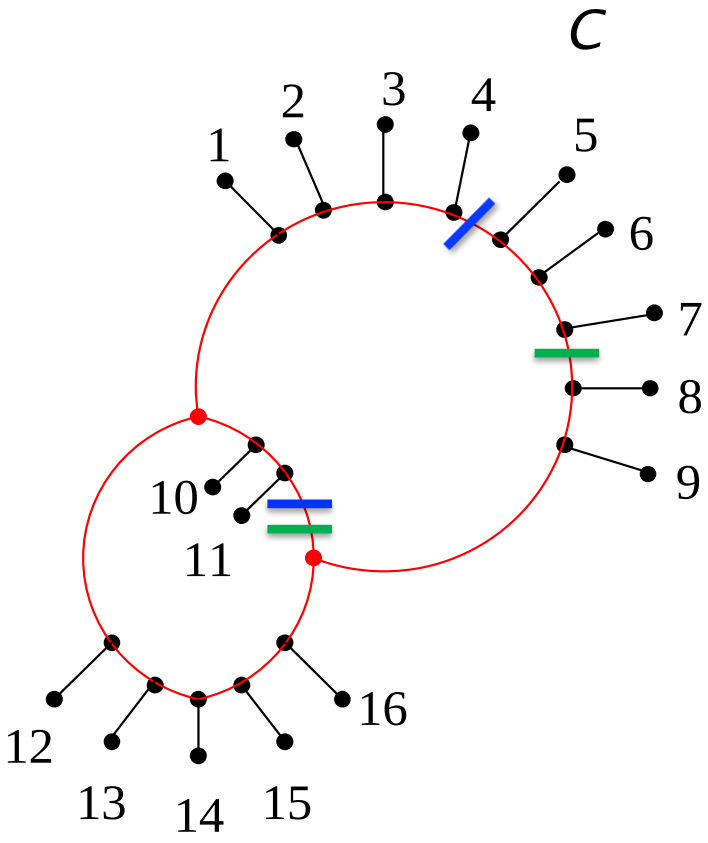


T'



C survives in one of *T* and *T'* and is broken in the other tree.

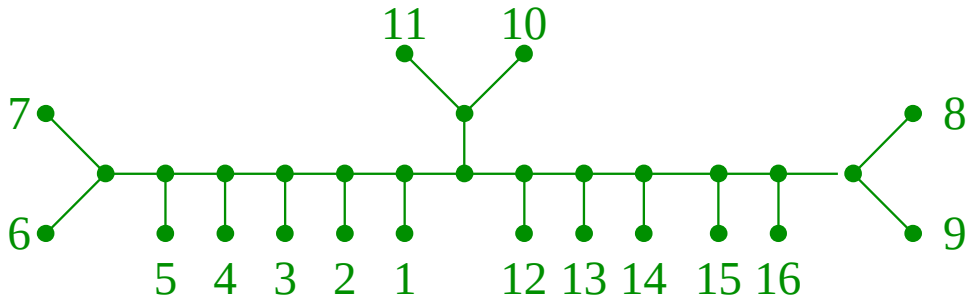
Breakpoints



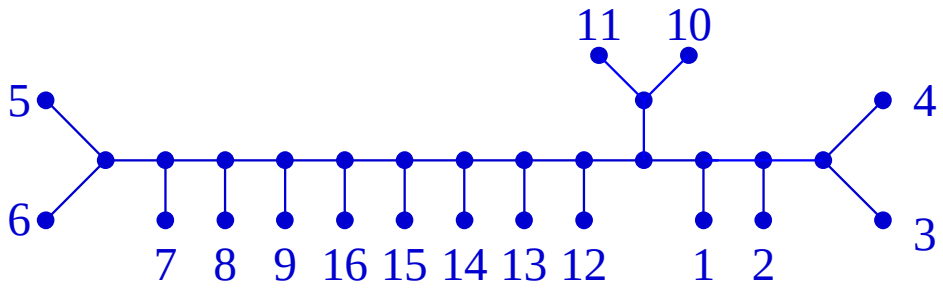
r(N)=2

Chain *C=(1,2,…,9)* has **two** breakpoints relative to *T* and *T'*.

T



T'



C is broken in *T* and *T'*.

Breakpoint Lemma. (K. and Linz, 2018).

Let S and S' be two trees with no common pendant subtree of size at least 2 and no common chain of length at least 4. Let N be a network that embeds S and S', and let C be an n-chain of N, where n is the length of C. Then

- $n \leq 3$ if C has no breakpoints relative to S and S',
- $n \leq 6$ if C has one breakpoint relative to S and S',
- $n \leq 9$ if C has two breakpoints relative to S and S'.

Breakpoint Lemma. (K. and Linz, 2018).

Let S and S' be two trees with no common pendant subtree of size at least 2 and no common chain of length at least 4. Let N be a network that embeds S and S', and let C be an n-chain of N, where n is the length of C. Then

- $n \leq 3$ if C has no breakpoints relative to S and S',
- $n \leq 6$ if C has one breakpoint relative to S and S',
- $n \leq 9$ if C has two breakpoints relative to S and S'.

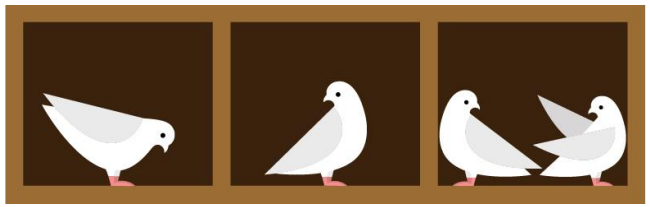
Can we have 10 or more leaves on a side? No!

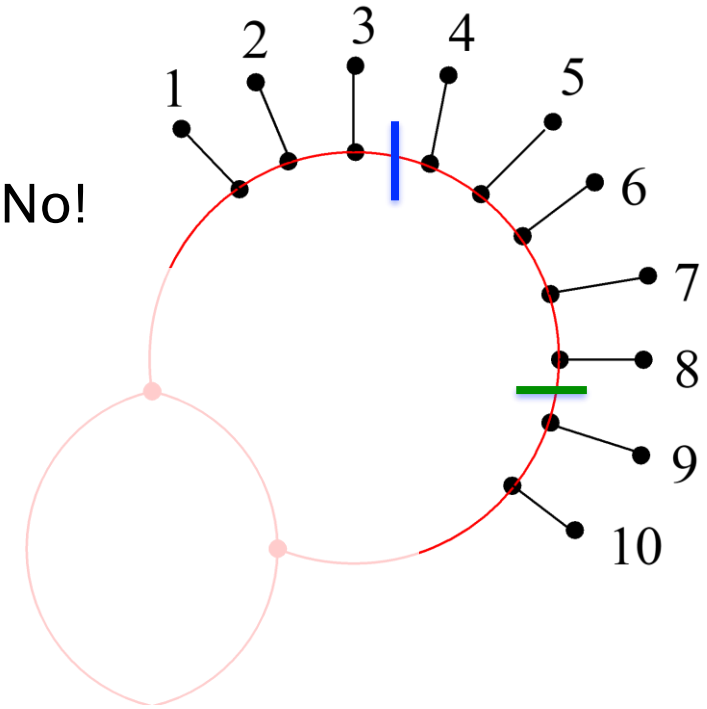
Breakpoint Lemma. (K. and Linz, 2018).

Let S and S' be two trees with no common pendant subtree of size at least 2 and no common chain of length at least 4. Let N be a network that embeds S and S', and let C be an n-chain of N, where n is the length of C. Then

- $n \leq 3$ if C has no breakpoints,
- $n \leq 6$ if C has one breakpoint,
- $n \leq 9$ if C has two breakpoints.

Can we have 10 or more leaves on a side? No!
Why? Apply pigeonhole principle.





Both trees have (4,5,6,7) as a chain!

Lemma. (K. and Linz, 2018).

Let S and S' be two trees on X' with no common pendant subtree of size at least 2 and no common chain of length at least 4. If $d_{\mathrm{TBR}}(S,S') \geq 2$, then

$$|X'| \leq 15 d_{\mathrm{TBR}}(S,S') - 9.$$

Lemma. (K, and Linz, 2018).

Let S and S' be two trees on X' with no common pendant subtree of size at least 2 and no common chain of length at least 4. If $d_{TBR}(S,S') \geq 2$, then

$$|X'| \leq 15 d_{TBR}(S,S') - 9.$$

Proof sketch. There are $2k$ breakpoints, to distribute across $3(k-1)$ sides.

Sides with 0, 1, 2 breakpoints can have at most 3, 6, 9 leaves respectively.

The optimum of the counting equation is $15k-9$.

Theorem. (Allen and Steel, 2001).

[Linear kernel] Let S and S' be two trees obtained from T and T' by repeated applications of the subtree and chain reduction until no further reduction is possible. Then

$$|X'| \leq 28 d_{\mathrm{TBR}}(T,T'),$$

where X' is the leaf set of S and S'.

Improved kernel is $|X'| \leq 15 d_{\mathrm{TBR}}(T,T')-9$. (K. and Linz, 2018)

Is the new kernel tight?

Theorem. (Allen and Steel, 2001).

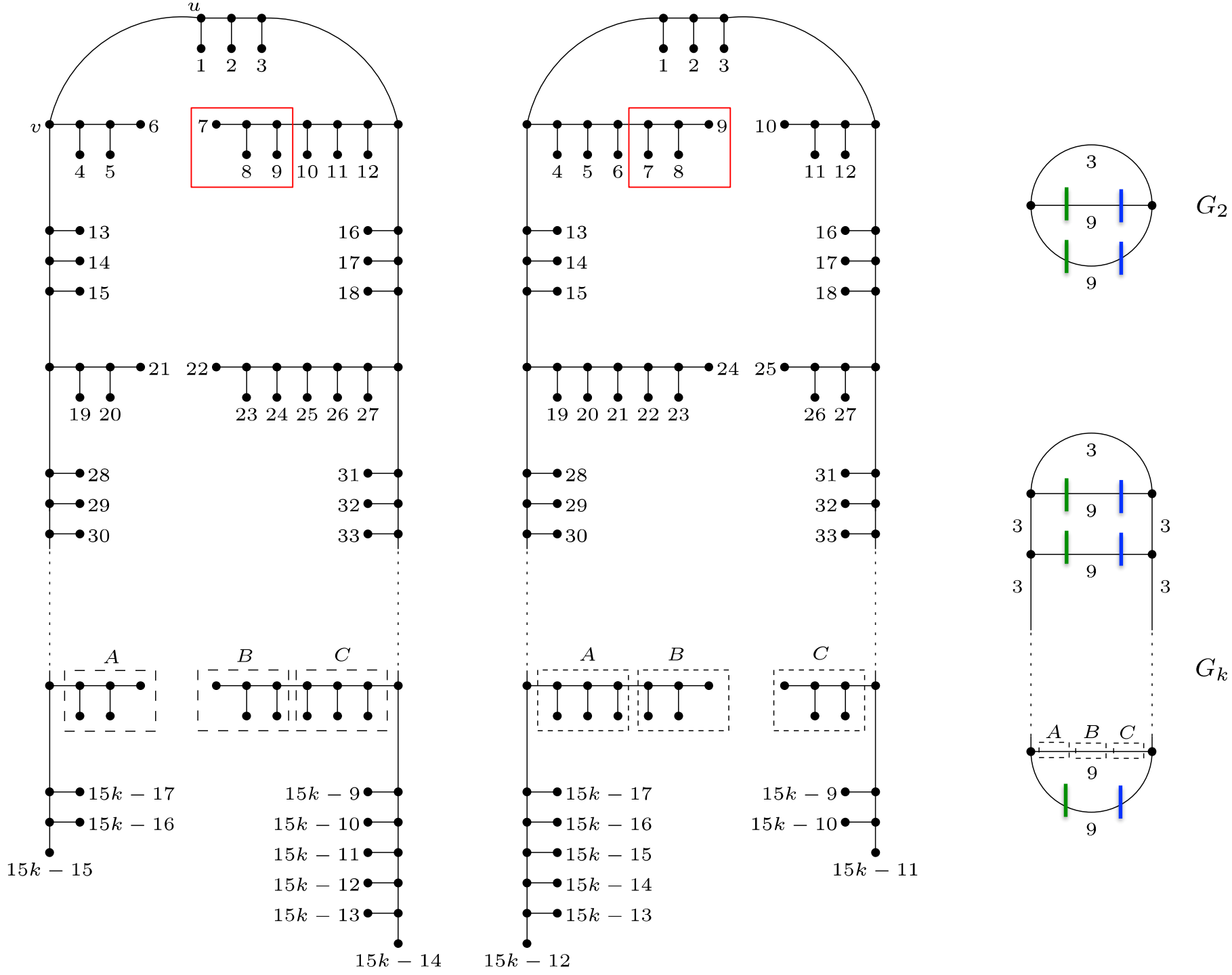
[Linear kernel] Let S and S' be two trees obtained from T and T' by repeated applications of the subtree and chain reduction until no further reduction is possible. Then

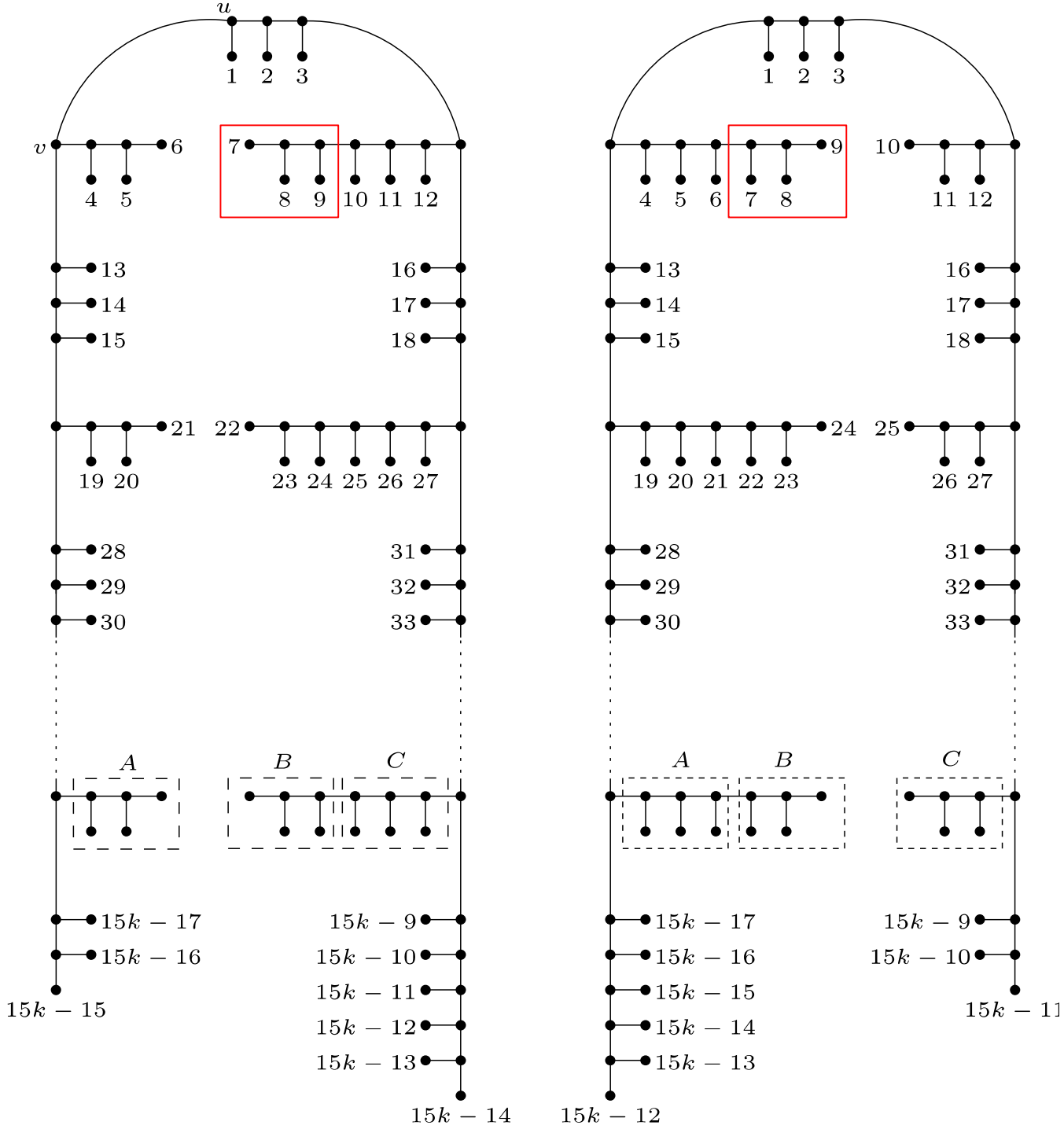
$$|X'| \leq 28 d_{\mathrm{TBR}}(T,T'),$$

where X' is the leaf set of S and S'.

Improved kernel is $|X'| \leq 15 d_{\mathrm{TBR}}(T,T')-9$. (K. and Linz, 2018)

Is the new kernel tight?
YES.


u
1 2 3
v
6
4 5
7
8 9 10 11 12
13
14
15
16
17
18
21
19 20
22
23 24 25 26 27
28
29
30
31
32
33
A
B
C
15k − 17
15k − 16
15k − 15
15k − 9
15k − 10
15k − 11
15k − 12
15k − 13
15k − 14
1 2 3
9
4 5 6 7 8
10
11 12
24
19 20 21 22 23
25
26 27
A
B
C
15k − 17
15k − 16
15k − 15
15k − 14
15k − 13
15k − 12
15k − 9
15k − 10
15k − 11
3
9
9
G_2
3
9
3
3
3
3
9
9
A
B
C
9
G_k

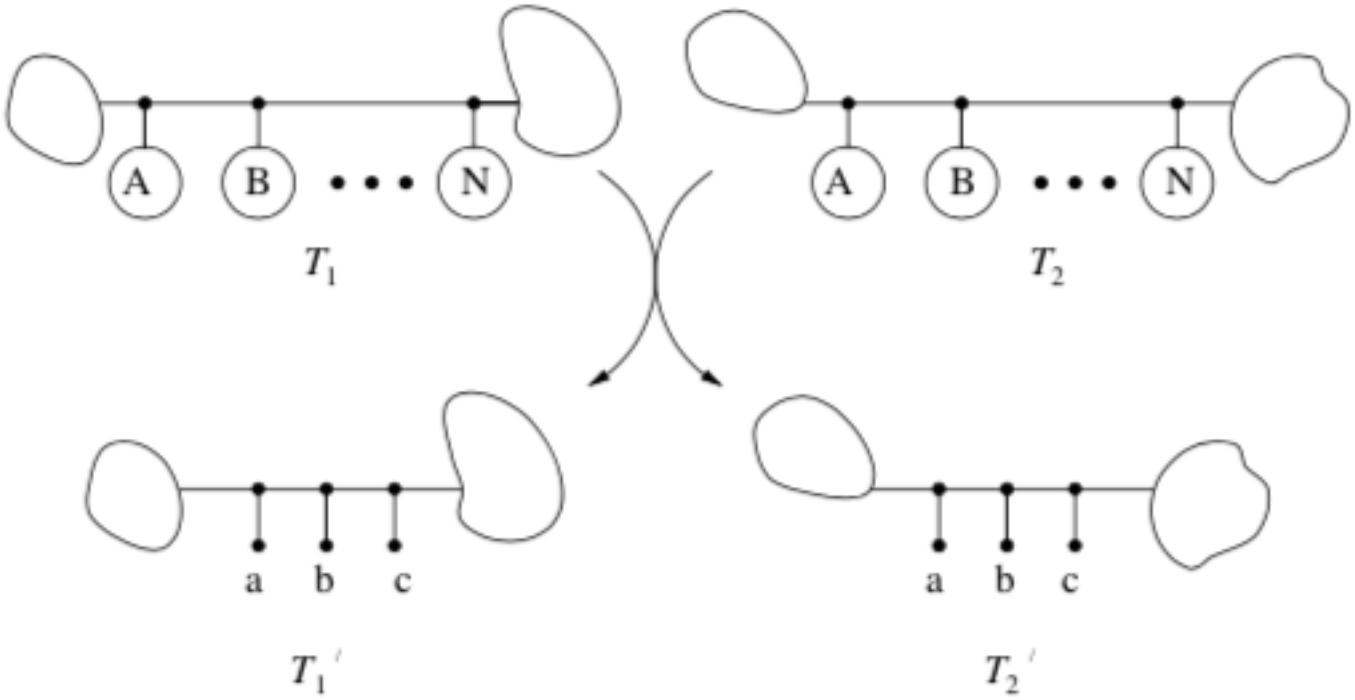


For each $k \geq 2$ we can build two trees such that:

- The reduction rules can no longer be applied;
- The TBR distance is exactly k;
- The two trees have exactly $15k-9$ leaves.

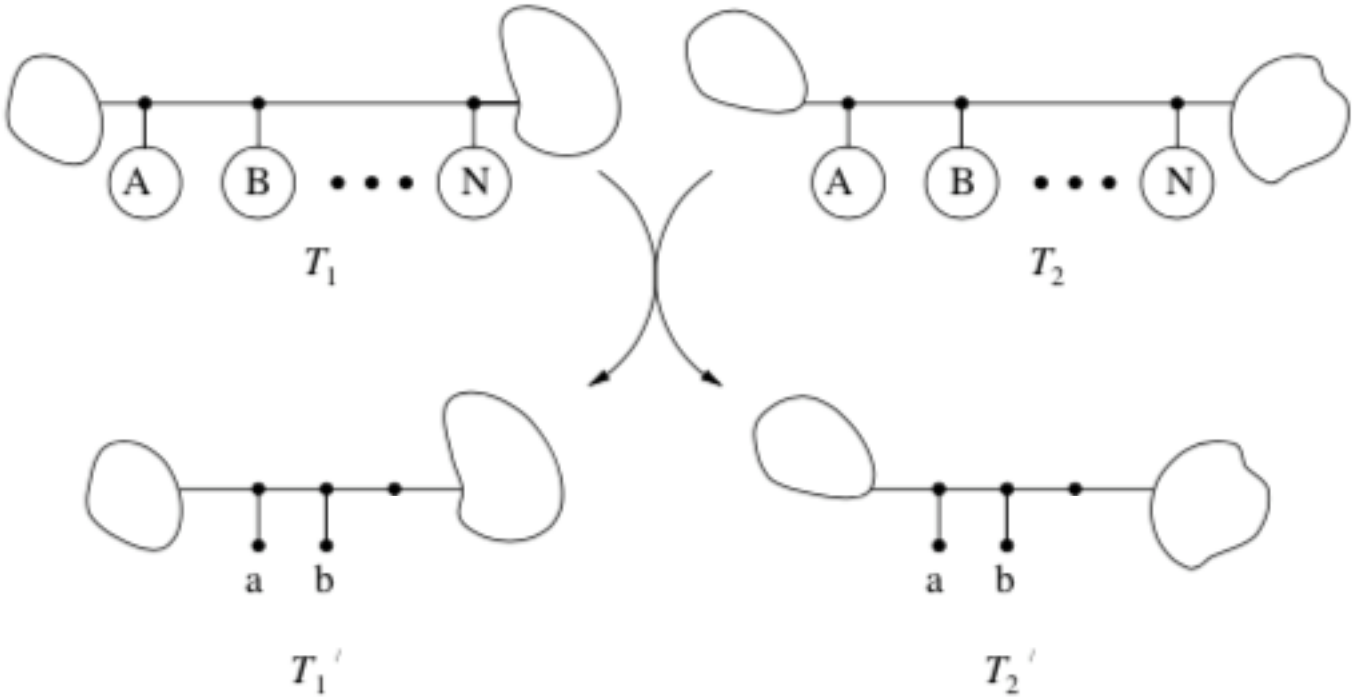
Can we do better?

Can we reduce chains to length 2?



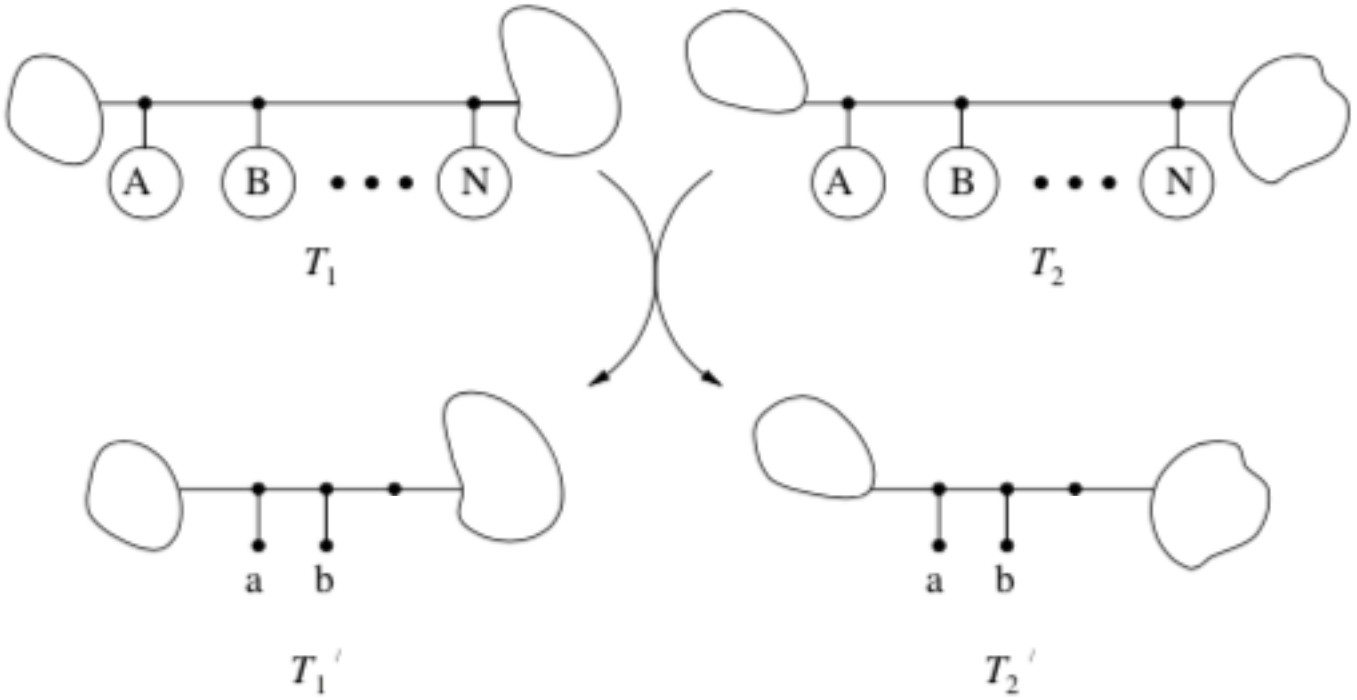
Allen and Steel, 2001

Can we reduce chains to length 2?



Allen and Steel, 2001

Can we reduce chains to length 2?



Allen and Steel, 2001

No! In some cases this causes d_{TBR} to decrease.

Can we do better? Yes!

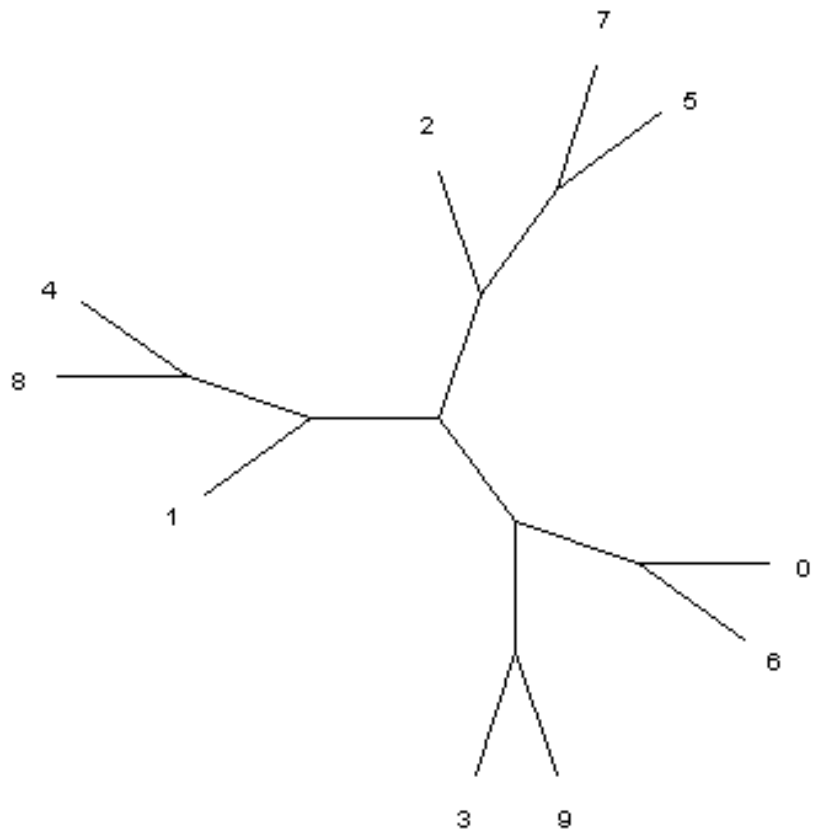
Idea. We describe 5 (!) new reduction rules which have been engineered to reduce the critical numbers in our counting argument:

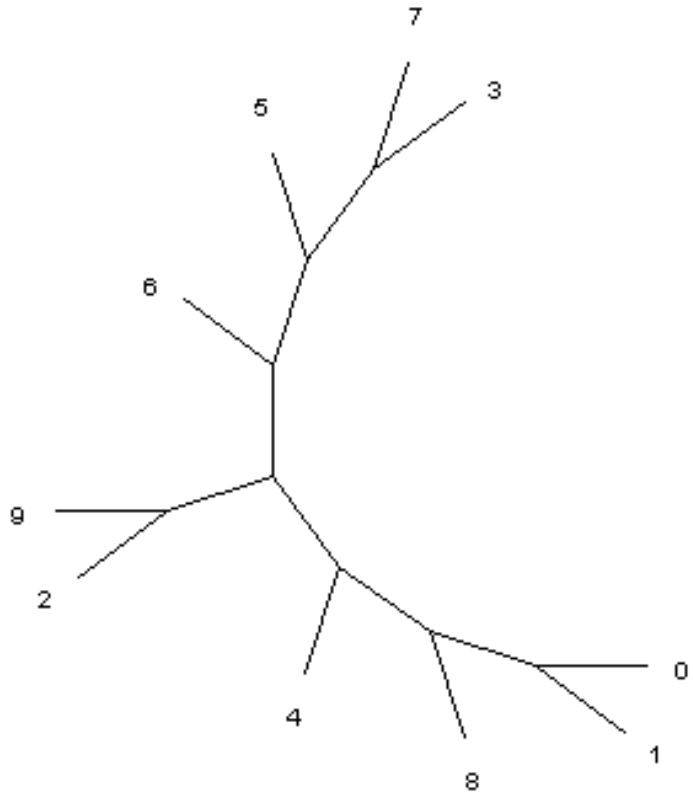
- $n \leq 3$ if C has no breakpoints,
- $n \leq 6 \rightarrow 4$ if C has one breakpoint,
- $n \leq 9 \rightarrow 4$ if C has two breakpoints.

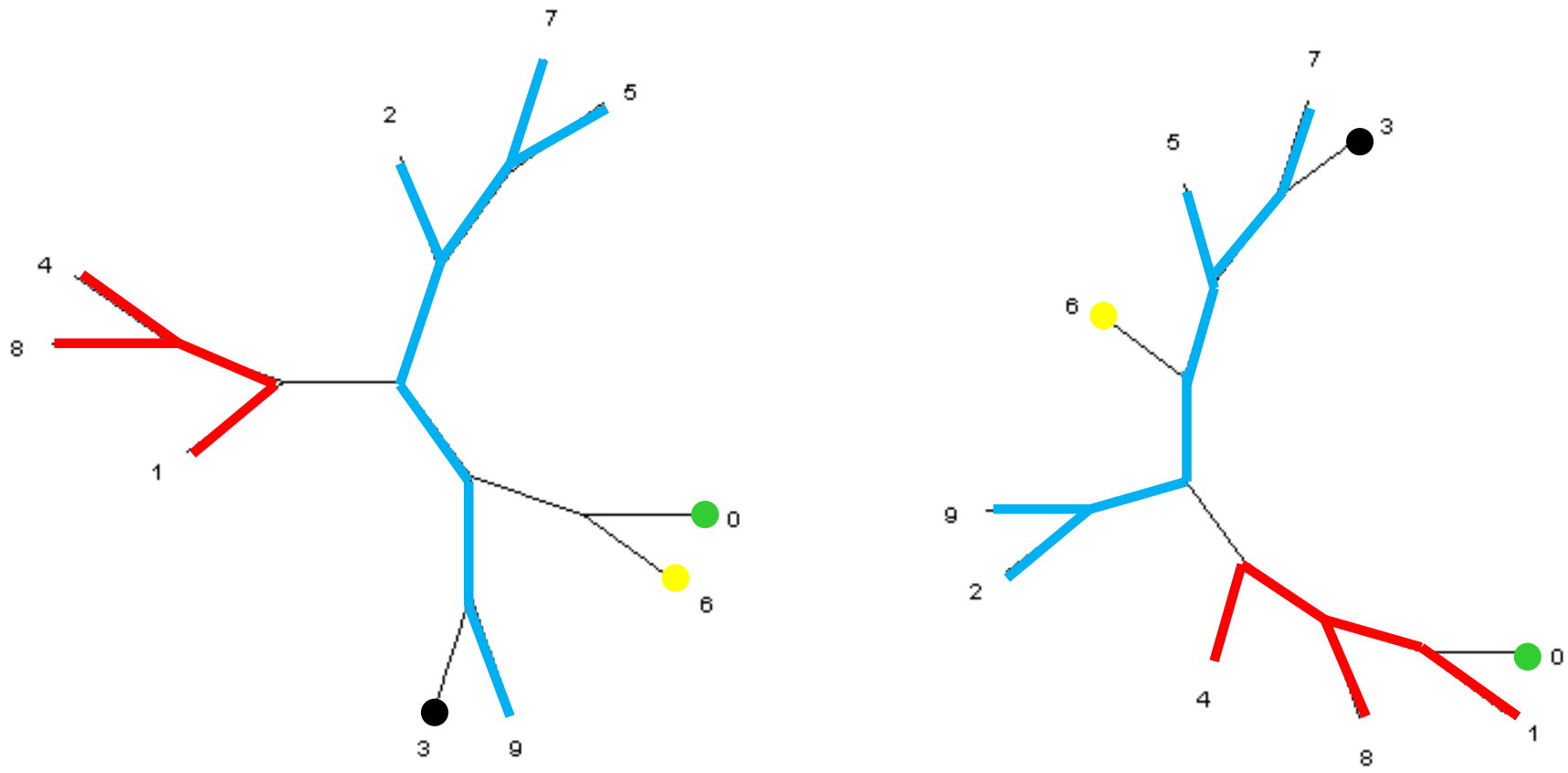
By dividing $2k$ breakpoints across $3(k-1)$ sides, we conclude that the size of the new kernel is at most…

$$4*2k + 3*(k-3) = 11k-9.$$

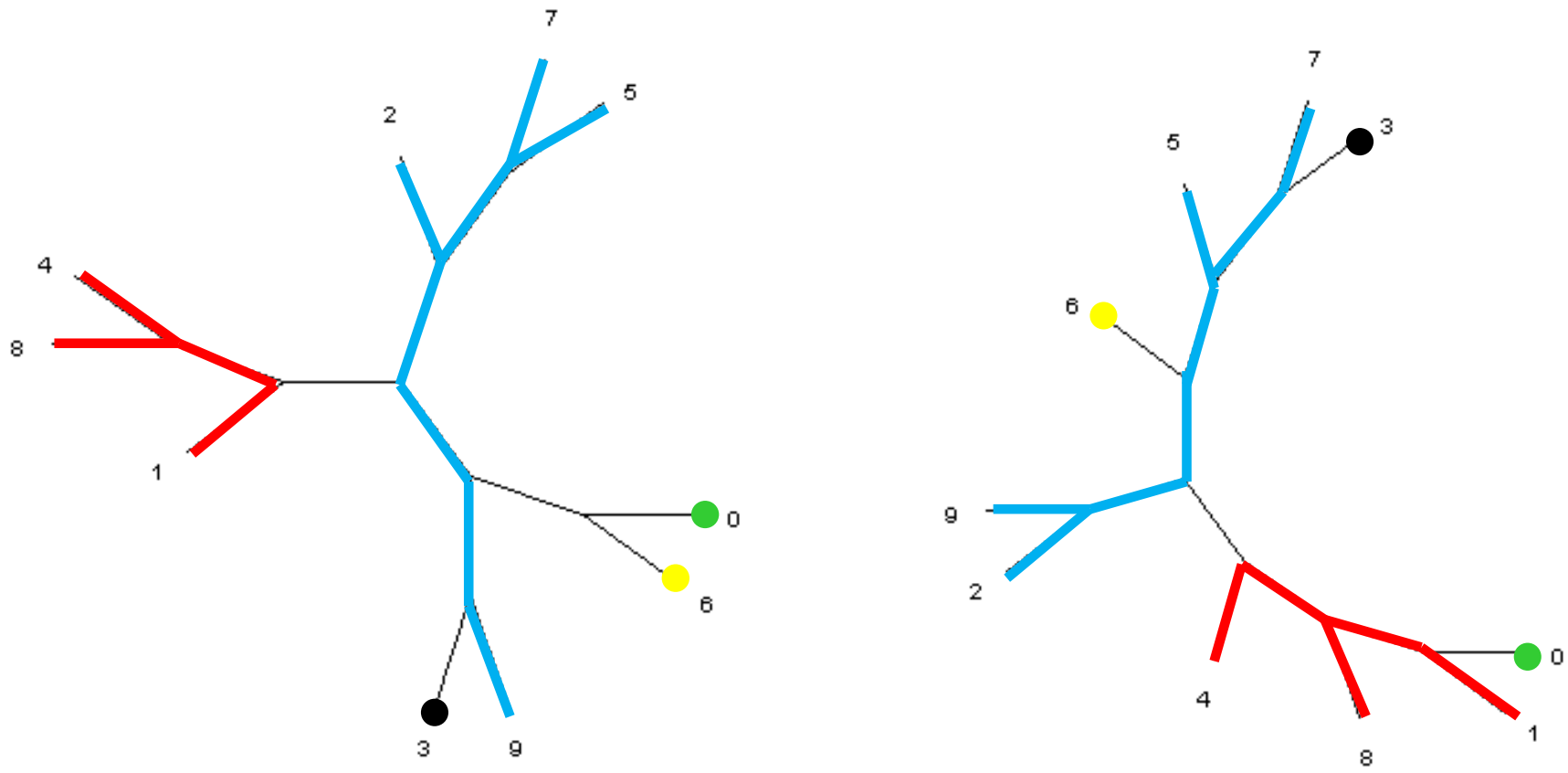
The correctness of these new rules requires use of the *agreement forest* abstraction.


7
5
2
4
8
1
0
6
3
9

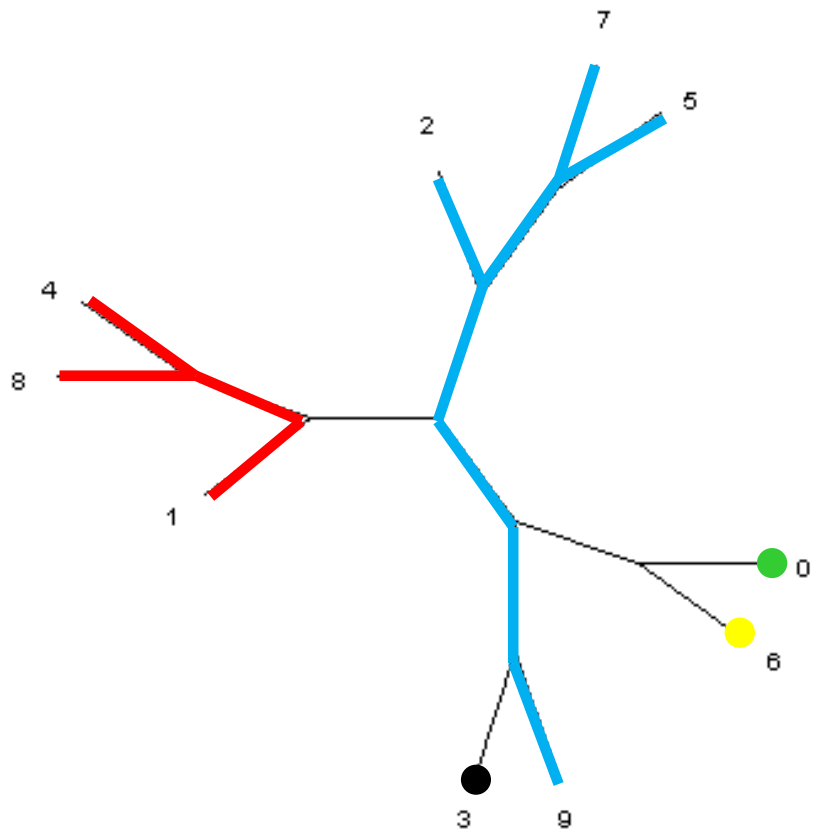

7
3
5
6
9
2
0
4
1
8

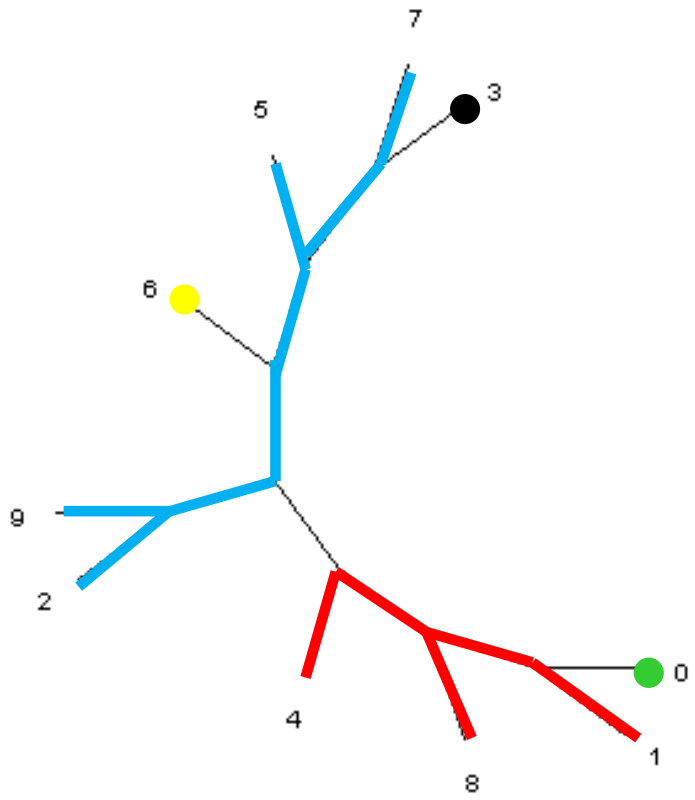


Agreement forest with 5 components



Fewer components are not possible: this is a *maximum agreement forest* (MAF)





Allen and Steel 2001:

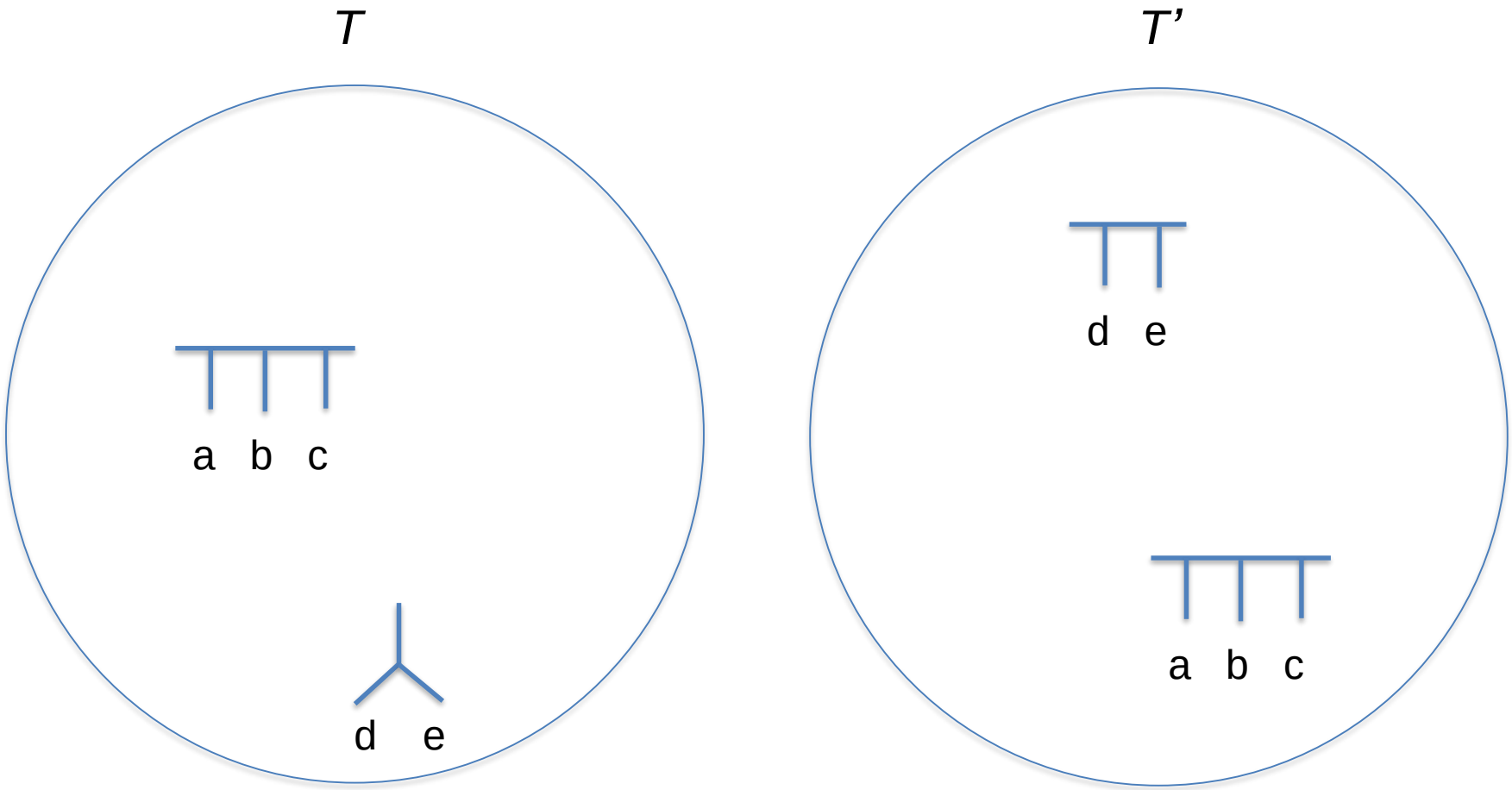
d_{TBR} is equal to the number of components in a MAF, minus 1.

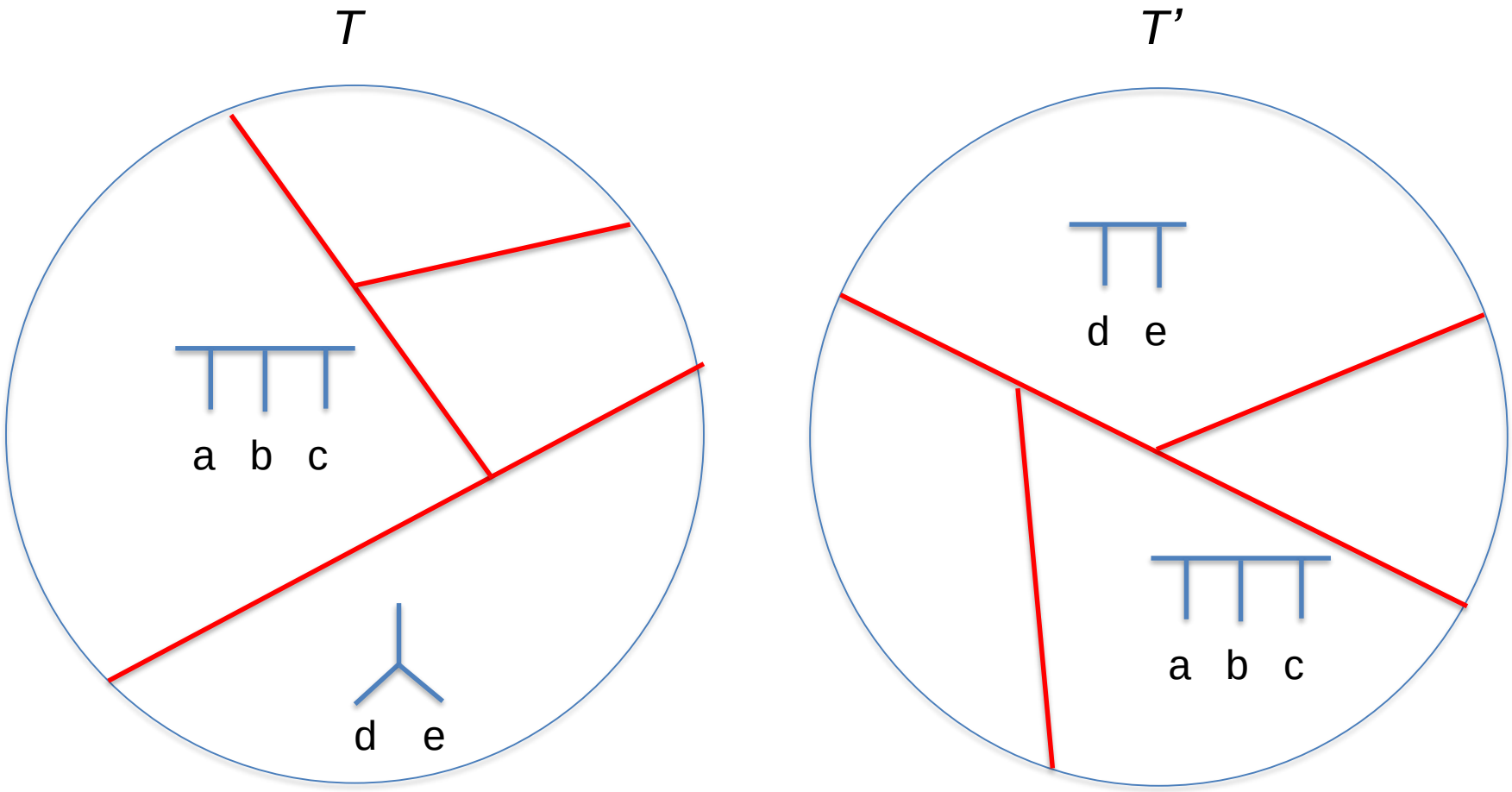
Chain preservation Theorem. (K. and Linz, 2019).

Let K be a set of disjoint common chains in T and T' , such that each chain in K has length ≥ 3, or has length 2 and is "pendant" in at least one of T and T'.

Then there exists a maximum agreement forest in which all the chains in K are *preserved*.

In other words: no chain in K is split across two or more components of the forest.


T
a b c
d e
T’
d e
a b c



Some *maximum agreement forest* has this structure.

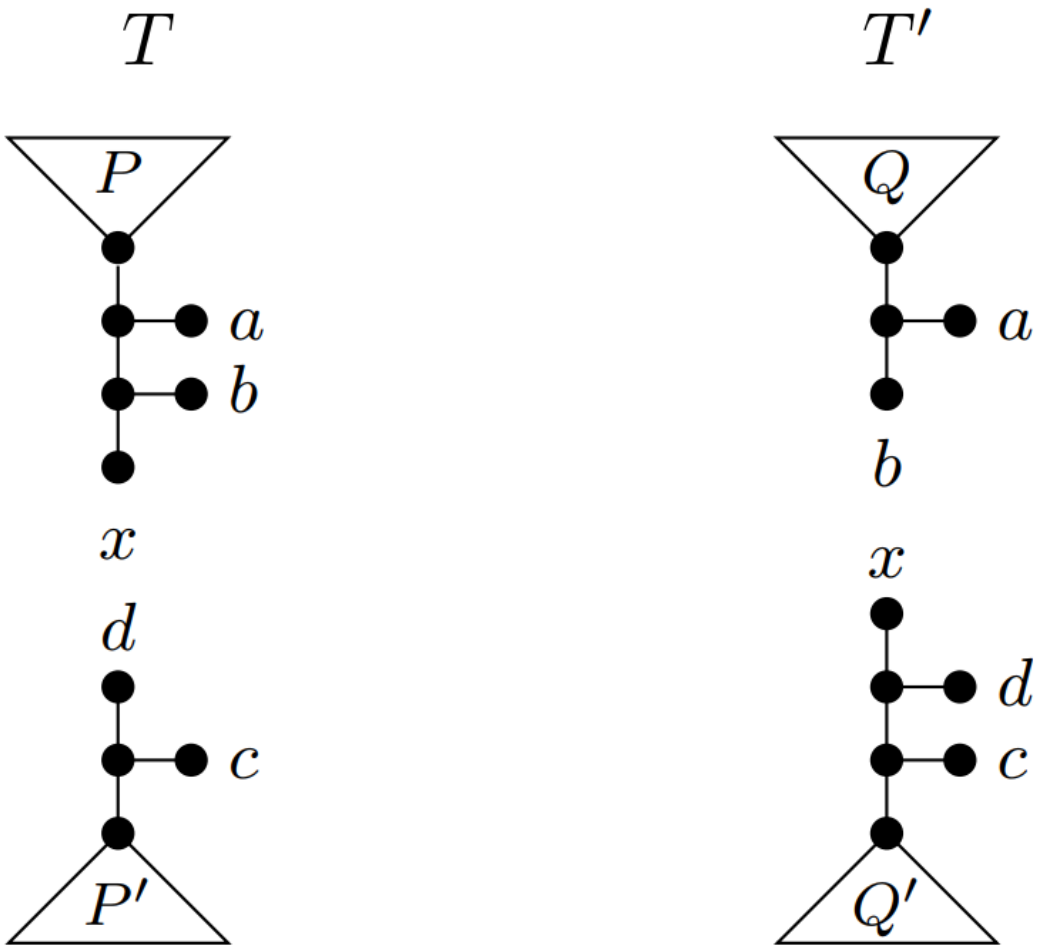
Using preserved chains as “obstructions”

Idea. Use the fact that chains are preserved to impose structure on some maximum agreement forest, such that at least one of the following holds:

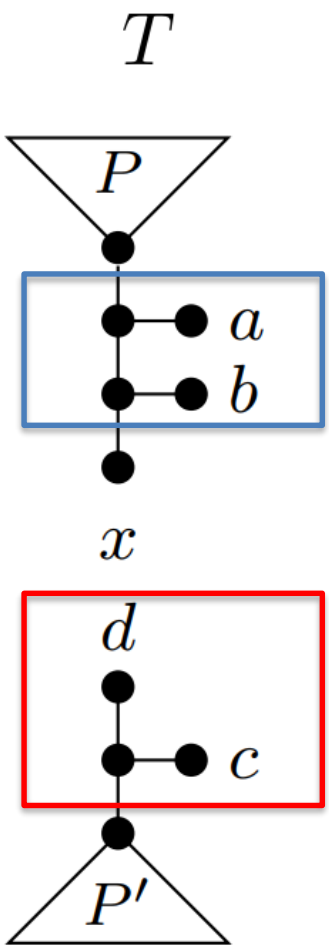
[Parameter reduction] Identify small subtrees whose deletion definitely reduces d_{TBR} by 1.

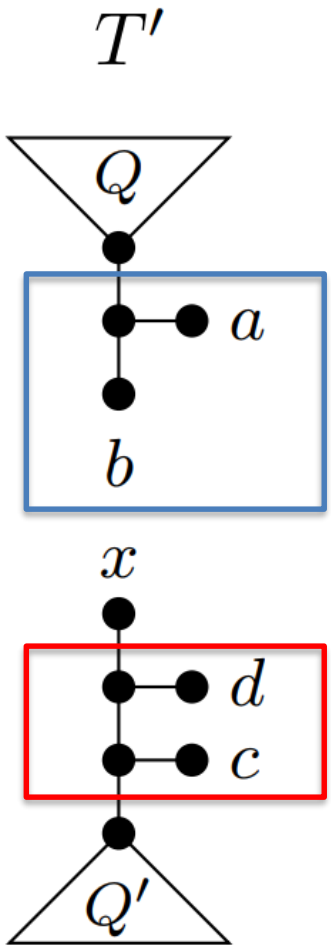
[Aggressive chain reduction] Identify short chains which can be reduced to length 2 (or even 1), without causing a decrease in d_{TBR}.

Example of parameter reduction



Example of parameter reduction

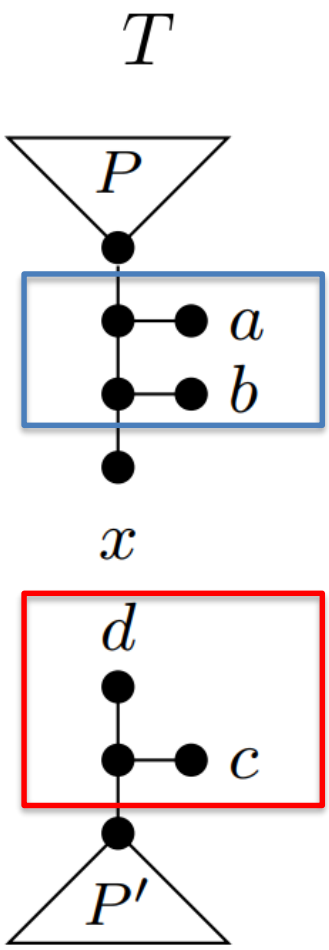


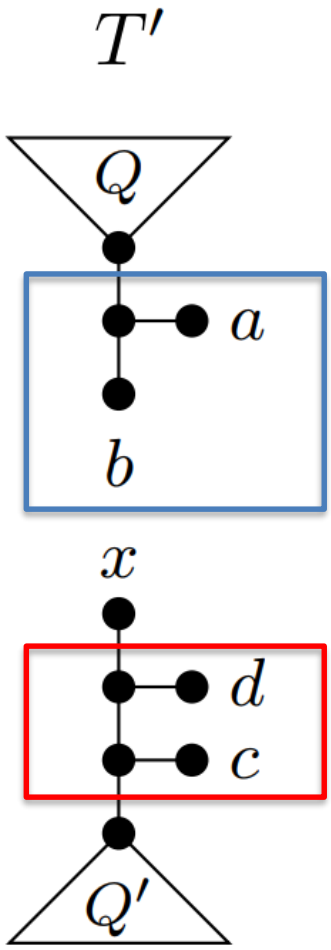


Common chain
{a,b}

Common chain
{c,d}

Example of parameter reduction



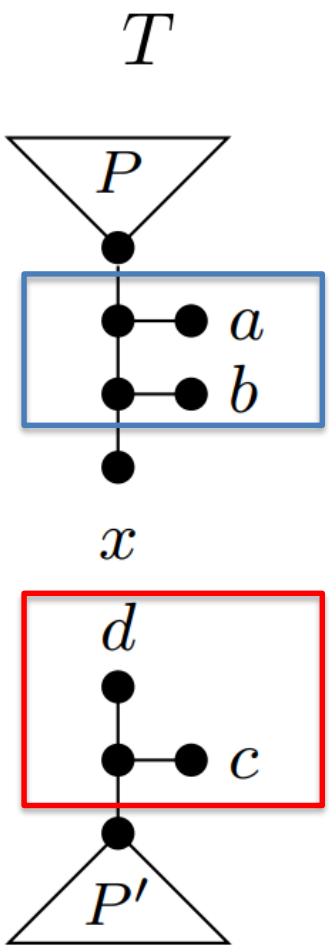


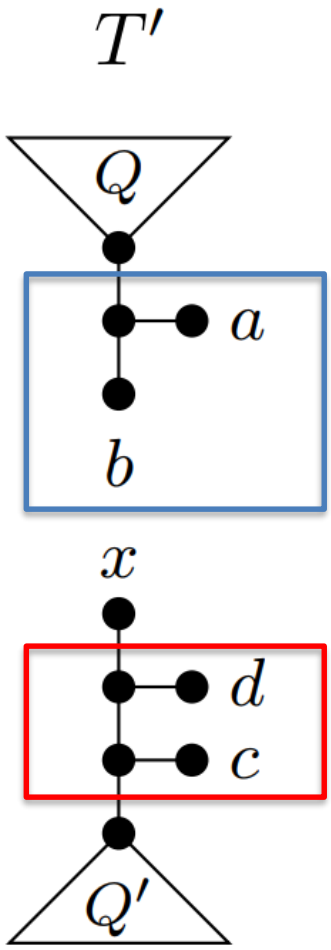
Common chain {a,b}

Common chain {c,d}

What about *x* … *?*

Example of parameter reduction



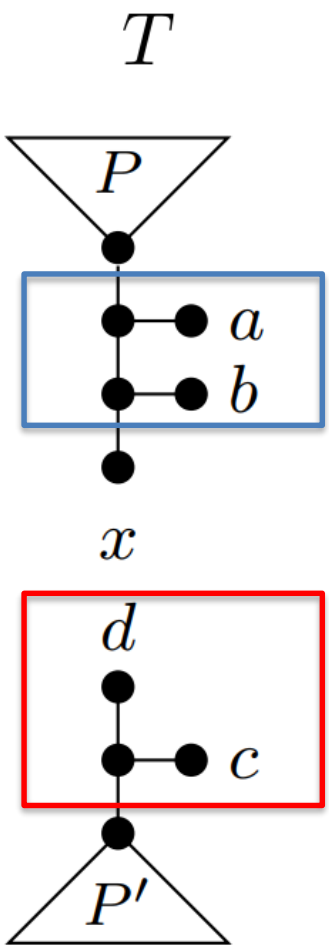


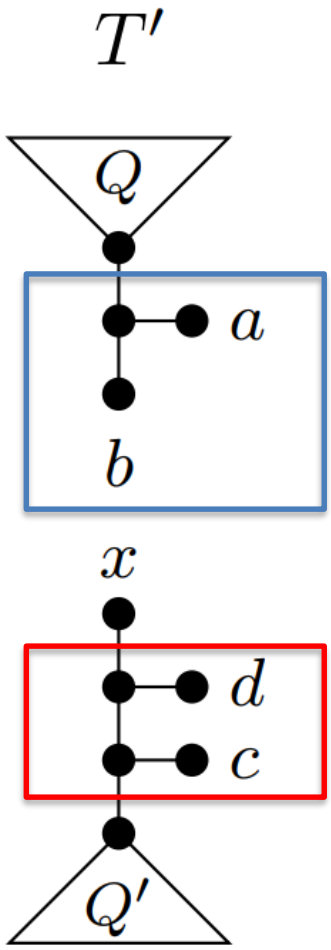
Common chain {a,b}

Common chain {c,d}

Can *x* be in the same component of the maximum agreement forest as {a,b}?

Example of parameter reduction



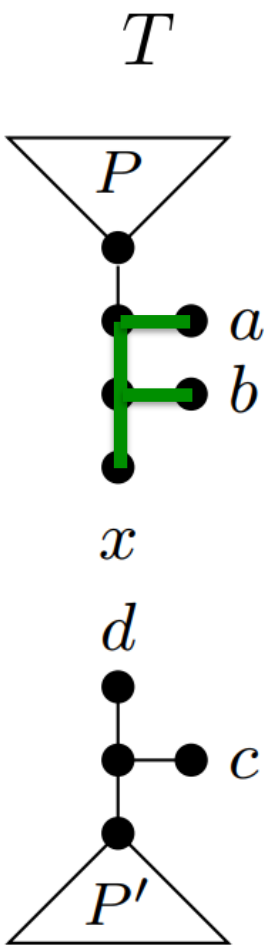


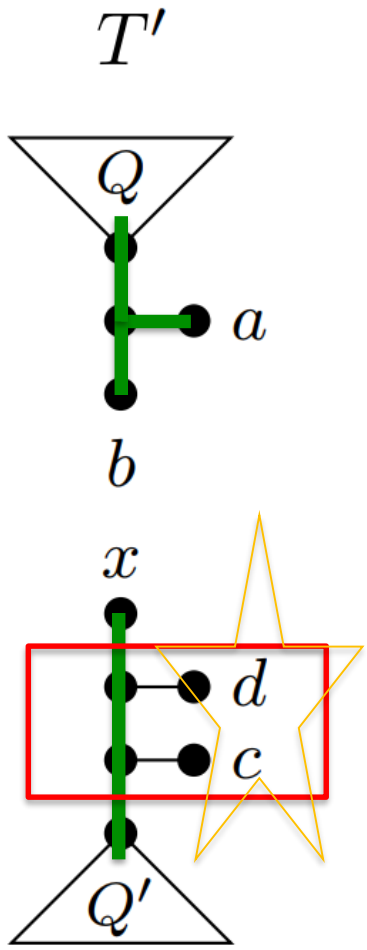
Common chain {a,b}

Common chain {c,d}

No! If {x,a,b} are together, this will cut through chain {c,d} in T', contradicting the preservation theorem

Example of parameter reduction





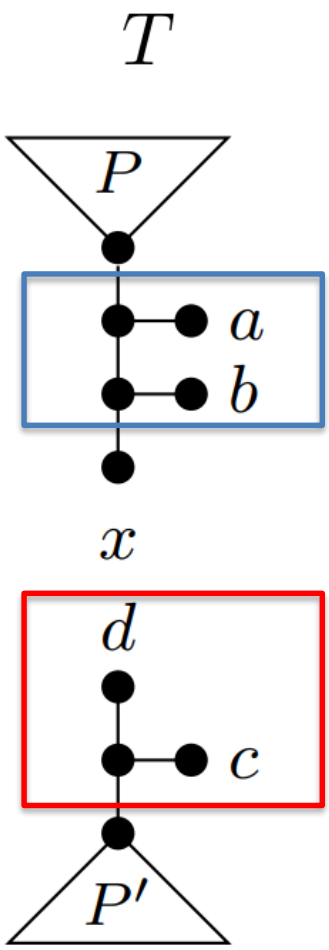
Common chain {a,b}

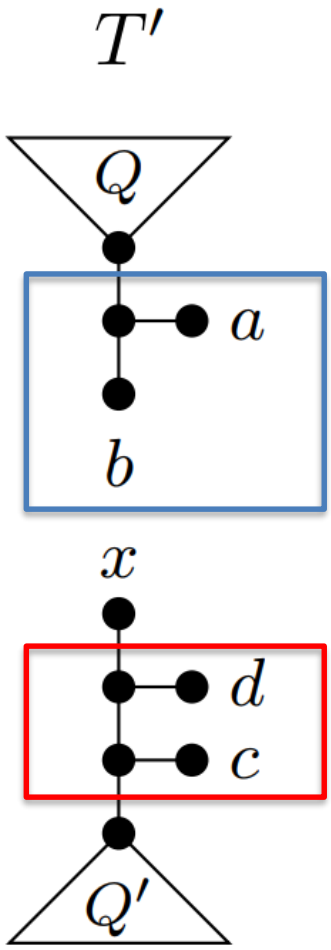
Shatters {c,d}!!

Common chain {c,d}

No! If {x,a,b} are together, this will cut through chain {c,d} in T', contradicting the preservation theorem

Example of parameter reduction



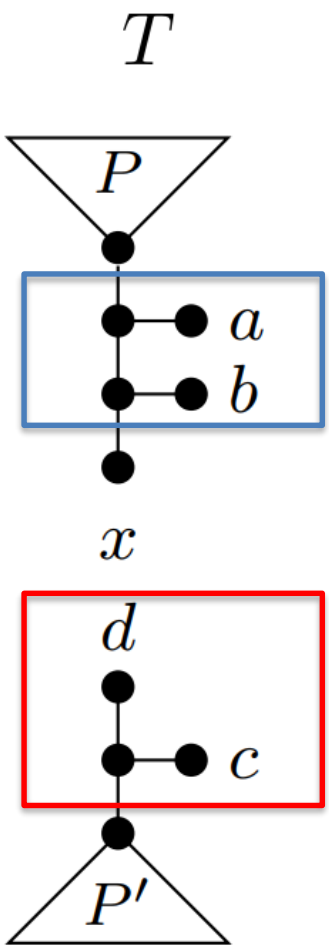


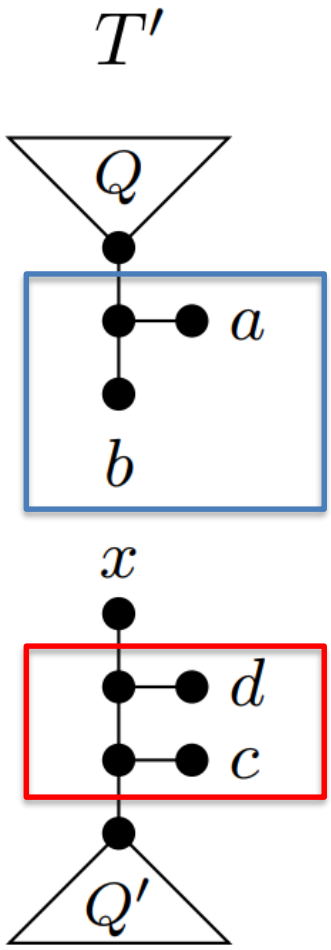
Common chain {a,b}

Common chain {c,d}

A symmetrical argument proves that *x* cannot join with {c,d}

Example of parameter reduction



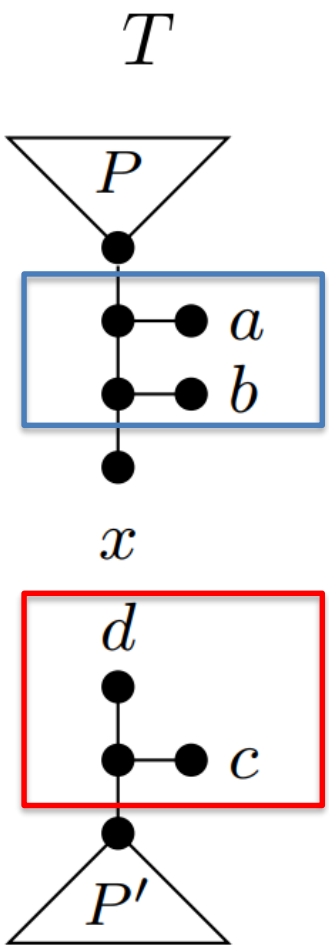


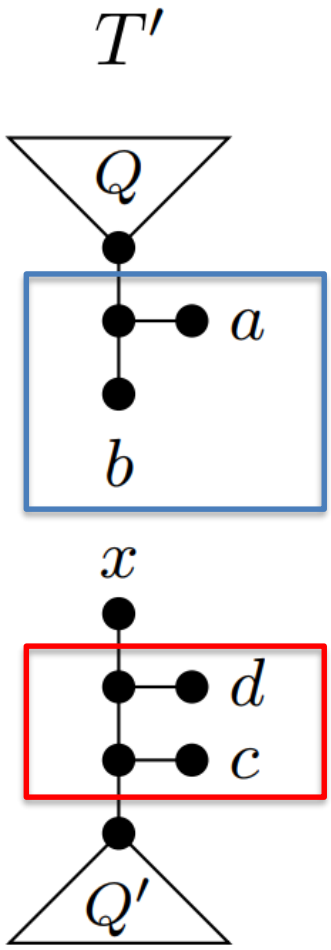
Common chain {a,b}

Common chain {c,d}

So *x* must be a singleton component of the maximum agreement forest

Example of parameter reduction



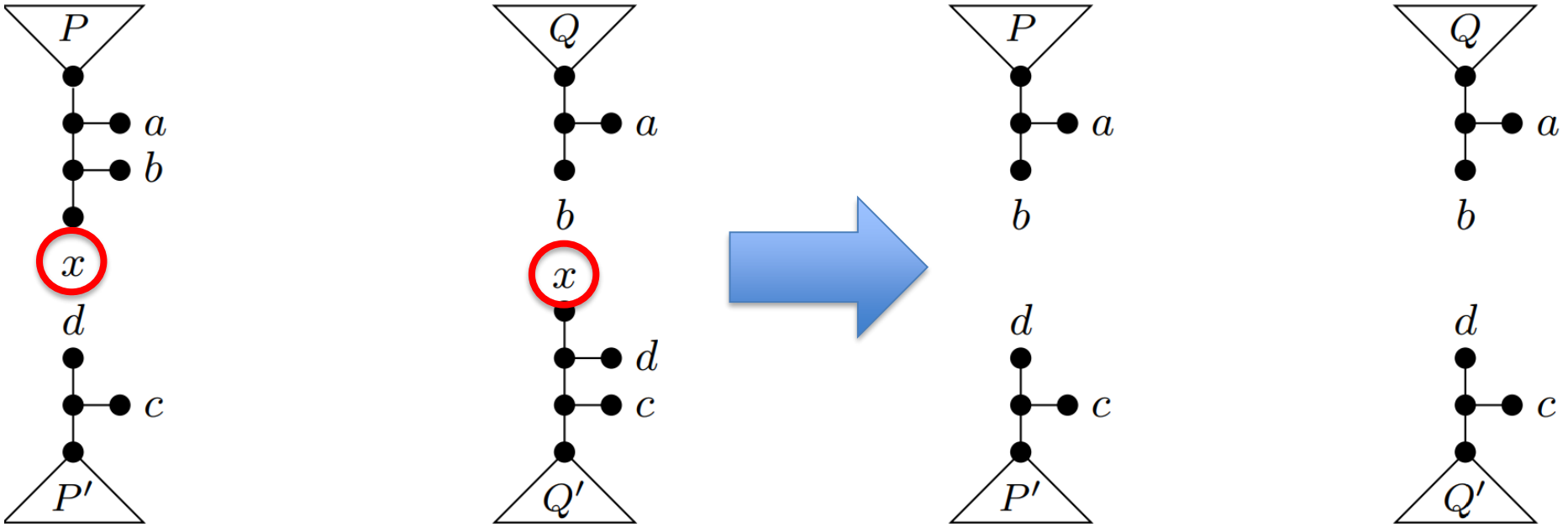


Common chain {a,b}

Common chain {c,d}

So deleting *x* from the trees must reduce dTBR by exactly one.

Example of parameter reduction

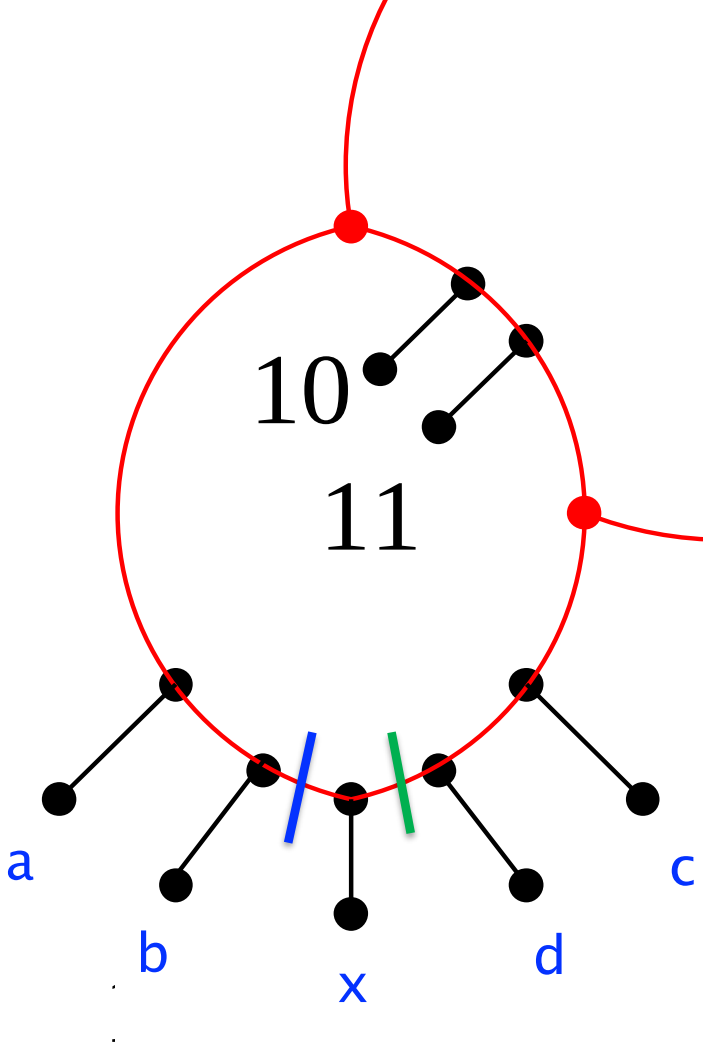


Original

T T'

Reduced

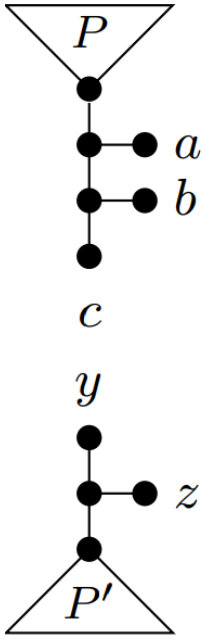
T T'

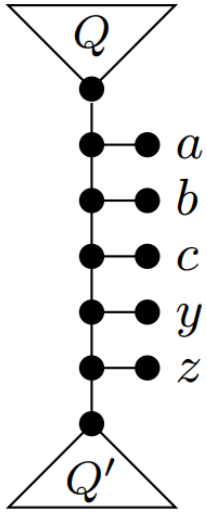


Idea. Assuming this reduction rule has been applied, you cannot have 5 taxa on a side of the network that has been divided by two breakpoints, as shown on the left.

(Many similar cases).

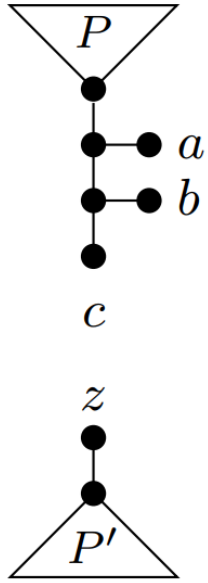
Finally: example of aggressive chain reduction (which *preserves* dTBR)

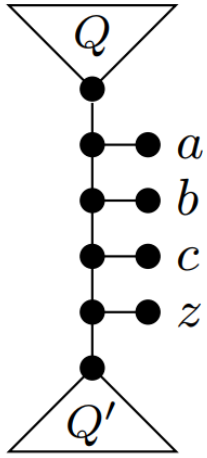




Original

T T'





Reduced

T T'

Putting it all together

In total we have introduced 5 new reduction rules which have been engineered to reduce the critical numbers in our counting argument:

- $n \leq 3$ if C has no breakpoints,
- $n \leq 6 \rightarrow 4$ if C has one breakpoint,
- $n \leq 9 \rightarrow 4$ if C has two breakpoints.

By dividing $2k$ breakpoints across $3(k-1)$ sides, we conclude that the size of the new kernel is at most…

$$4*2k + 3*(k-3) = 11k-9.$$

Moreover, this bound is (again) tight. (K. and Linz, 2019).

Conclusions and future work

Our new kernel for d_{TBR} has been achieved by simultaneously analysing the same problem from both a *graph-construction* and *agreement forest* perspective.

As far as we know, these are the first reduction rules to strictly enhance the reductive power of the classical subtree and chain reduction rules!

Can we go below *11k-9* ? Yes, we think so, but it will require new techniques (work in progress…) What is the limit?

Can we leverage these results to enhance FPT branching algorithms for computation of TBR distance?

In how far can we adapt the technique to work for other phylogenetic distances?

Thank you!

More details at:

- **A tight kernel for computing the tree bisection and reconnection distance between two phylogenetic trees,** https://arxiv.org/abs/1811.06892 (K. and Linz 2018)
- **New reduction rules for the tree bisection and reconnection distance,** https://arxiv.org/abs/1905.01468 (K. and Linz 2019)