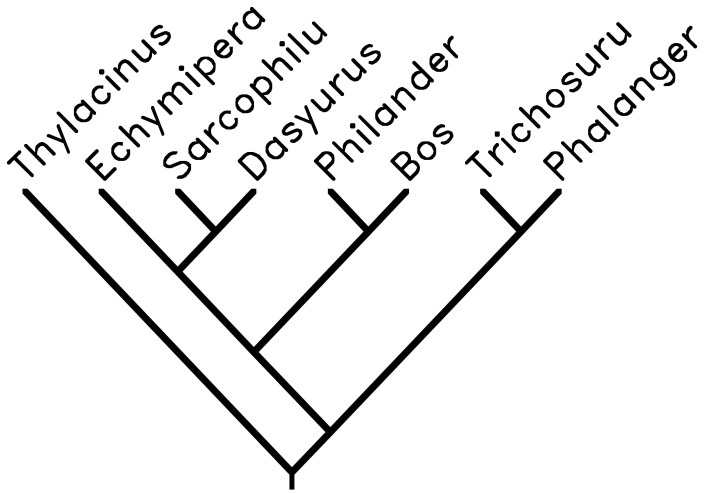
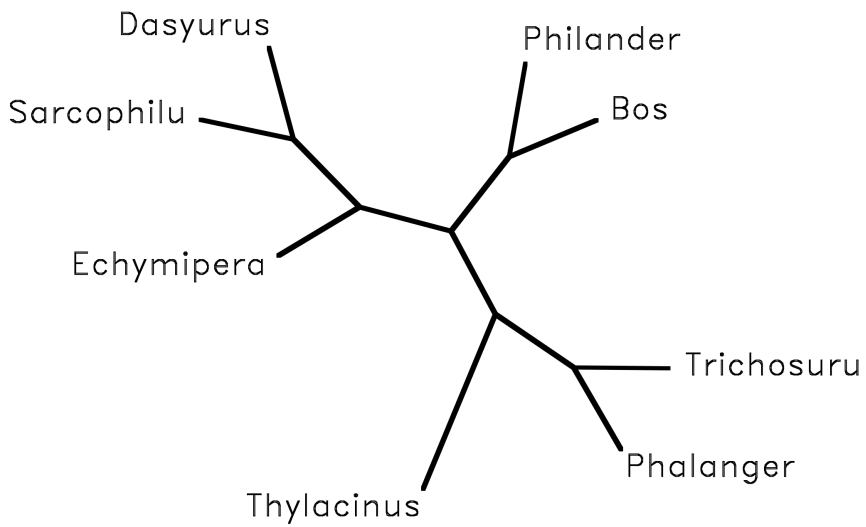
Swinging from Tree to Tree: Rearrangement Operations and their Metrics

Stefan Grünewald

CAS-MPG Partner Institute for Computational Biology
Shanghai, China

Phylogenetic Trees



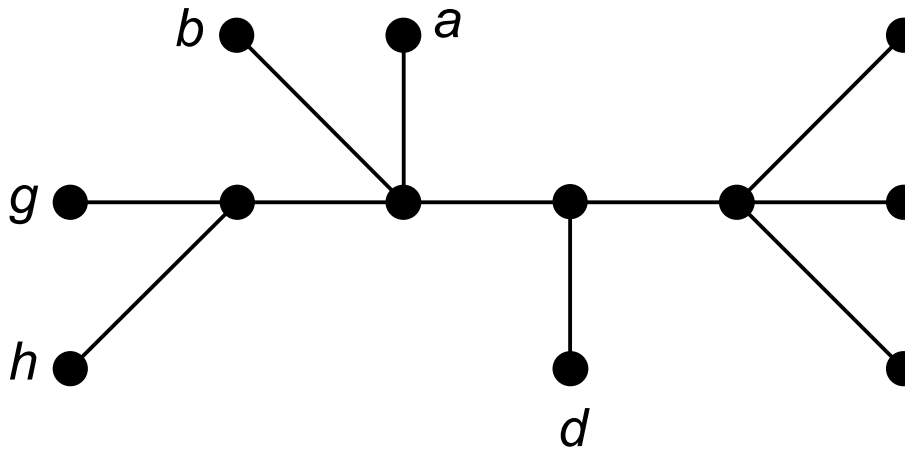


Phylogenetic Trees

A *phylogenetic tree T* is a (graph theoretic) tree without vertices of degree 2.

Its leaf set *L(T)* is also called the *taxa set.*

T is called *binary* if all interior vertices have degree 3.

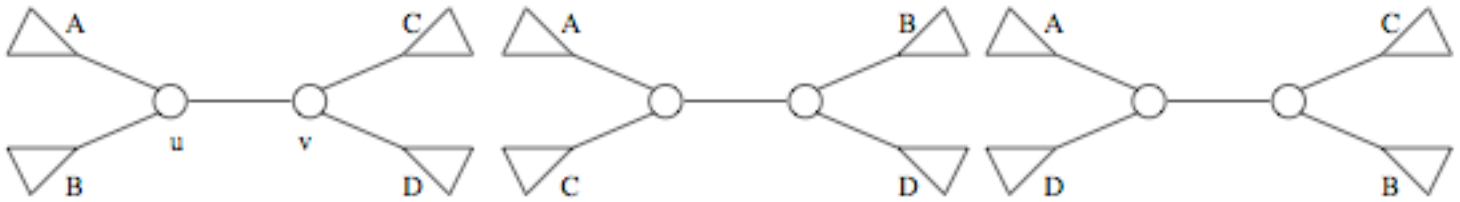


Different trees with the same taxa

- In phylogenetics one often observes several different trees on the same taxa set, e.g. by using different methods or analyzing different genes.
- Therefore, it is important to quantify how different two trees are.
- One common way to do so is using tree rearrangement operations

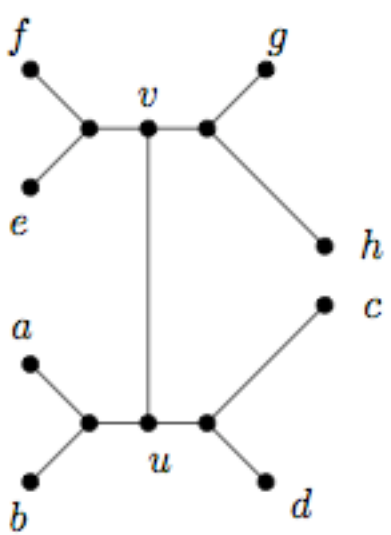
Nearest Neighbour Interchange (NNI)

- An NNI operation on an unrooted binary phylogenetic tree consists of identifying the two vertices *u* and *v* incident with an internal edge and then resolving it in one of the two different ways.

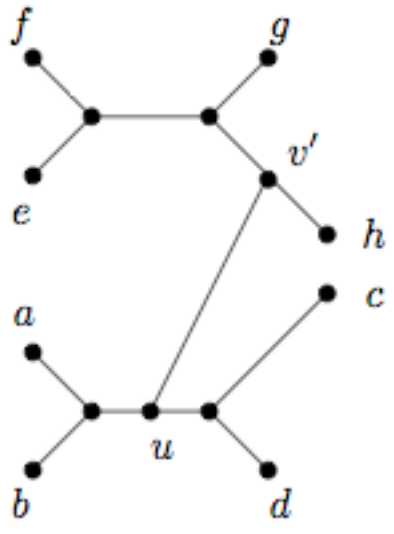


Bigger steps

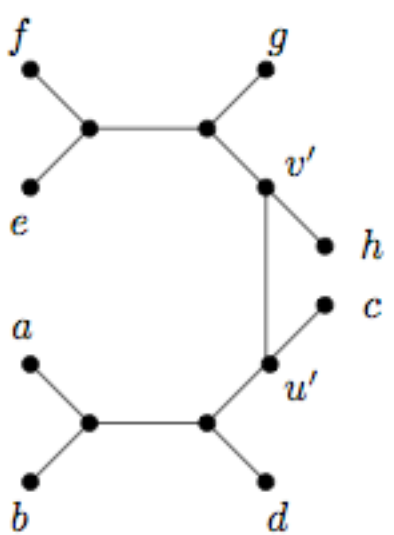
Subtree Prune and Regraft (SPR) and
Tree Bisection and Reconnection (TBR) operations consit of removing an interal edge and connecting the 2 resulting components differently.



T_1



T_2



T_3

SPR Operations

An SPR operation on an unrooted phylogenetic X-tree T is defined as follows:

- Remove an edge uv from T such that the component that contains v contains at least three taxa.
- Choose an edge that is not incident with v from the component of $T - uv$ that contains v and subdivide it by a new vertex w.
- Insert an edge uv.
- Suppress the vertex v of degree 2.

Distances

- Let $\boldsymbol{T}_n$ be the set of all phylogenetic trees with taxa set $\{1,\ldots,n\}$.
- For $\Theta \in \{\text{NNI, SPR, TBR}\}$, let $G(n, \Theta)$ be the graph with vertex set $\boldsymbol{T}_n$ where two vertices are adjacent, if one can be obtained from the other by performing a Θ-operation.
- The graph distance of $G(n, \Theta)$ defines a distance d_Θ on $\boldsymbol{T}_n$.

Applications of SPR

- The SPR distance has been used to estimate the amount of lateral gene transfer.
- SPR moves are used to escape from local optima in (meta-)heuristics to construct phylogenetic trees.

Unit neighborhood

- The size of the neighborhood of a tree with n taxa (the degree of a vertex in $G(n, \Theta)$)
- equals $2(n\text{-}3)$ for NNI
- equals $2(n\text{-}3)(2n\text{-}7)$ for SPR (Allen, Steel, 2001),
- depends on the tree shape and there are a lower bound $O(n^2 \log n)$ and an upper bound $O(n^3)$ for TBR (Humphries, Wu, preprint).

The diameter

- The diameter of G(n,NNI) is known and $O(n \log n)$, Li et al. 1996.
- The diameter of G(n,SPR) is between $1/2\ n - o(n)$ and n-3, Allen and Steel 2001.
- The diameter of G(n,TBR) is between $1/4\ n - o(n)$ and n-3, Allen and Steel 2001.

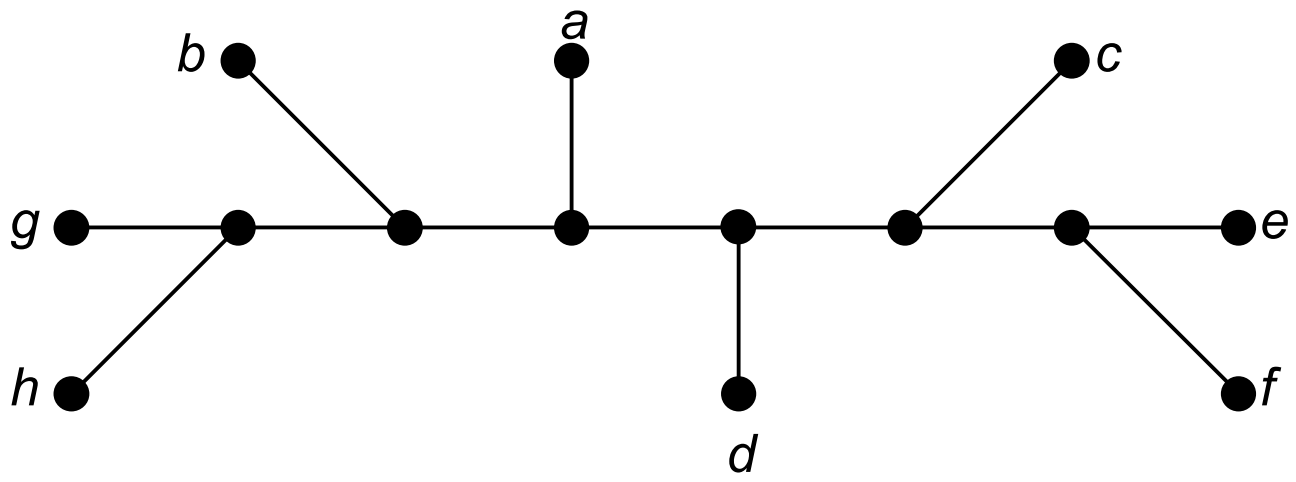
The diameter

- The diameter of G(n,NNI) is known and $O(n \log n)$, Li et al. 1996.
- The diameter of G(n,SPR) is between $1/2\ n - o(n)$ and n-3, Allen and Steel 2001.
- The diameter of G(n,TBR) is between $1/4\ n - o(n)$ and n-3, Allen and Steel 2001.
- **Theorem** (Ding, SG, Humphries, submitted):

$$n - 2\left\lceil \sqrt{n} \right\rceil + 1 \le \Delta_{\mathrm{TBR}}(n) \le \Delta_{\mathrm{SPR}}(n) \le n - \left\lfloor \tfrac{1}{2}\sqrt{n} \right\rfloor$$

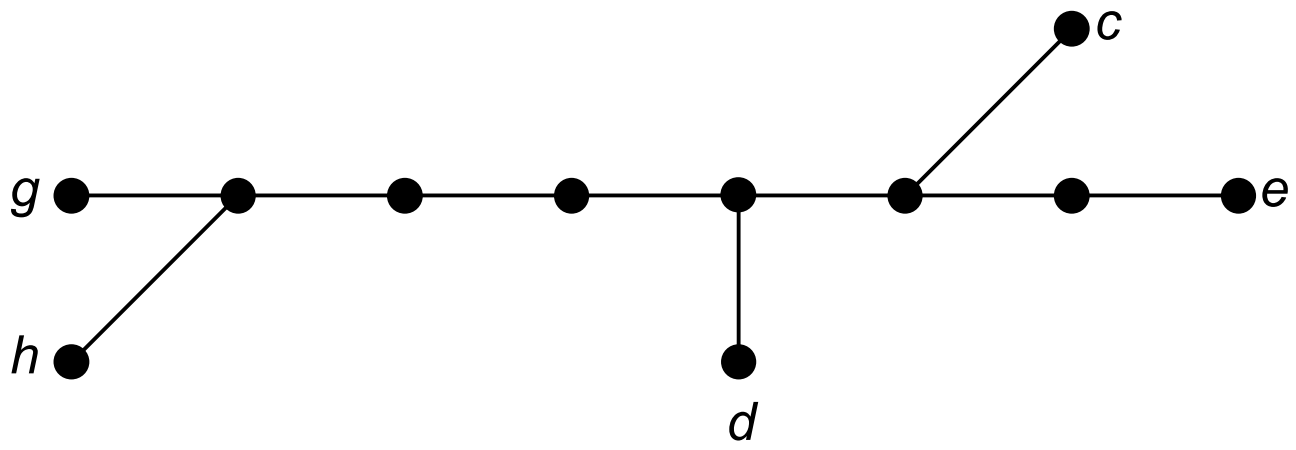
Restrictions

A *restriction* of a phylogenetic tree T to a subset S of $L(T)$ is the tree obtained from the smallest subtree of T containing S by suppressing all vertices of degree 2.



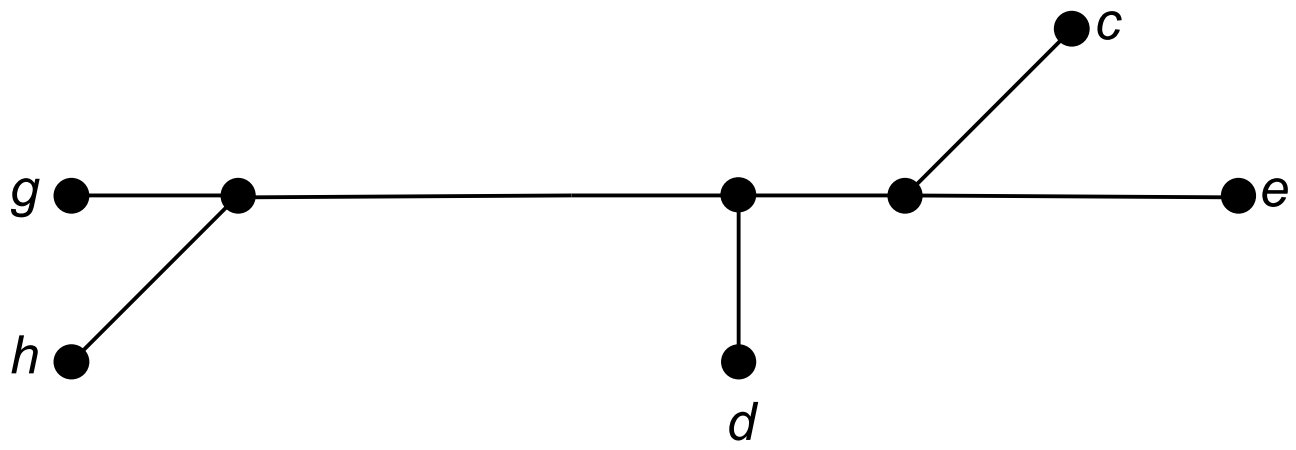
Restrictions

A *restriction* of a phylogenetic tree T to a subset S of $L(T)$ is the tree obtained from the smallest subtree of T containing S by suppressing all vertices of degree 2.



Restrictions

A *restriction* of a phylogenetic tree T to a subset S of $L(T)$ is the tree obtained from the smallest subtree of T containing S by suppressing all vertices of degree 2.



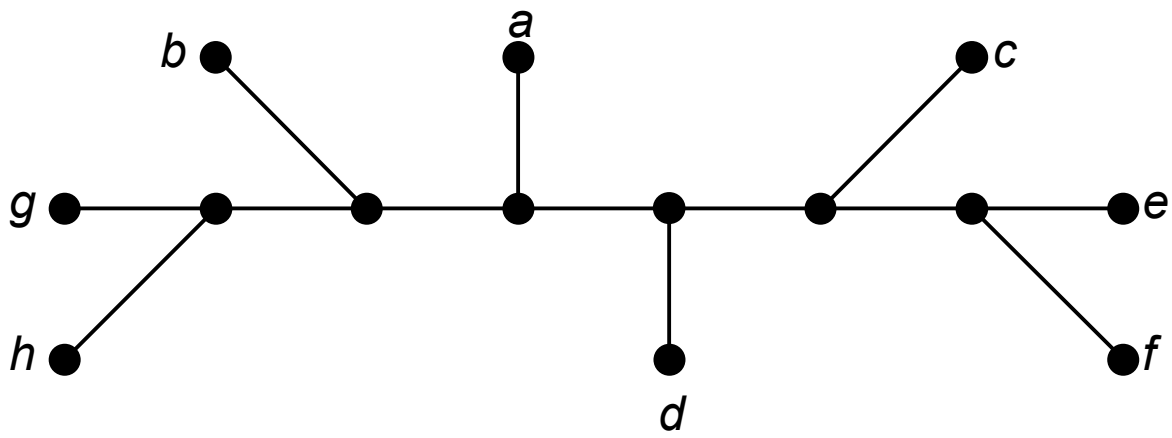
Agreement forests

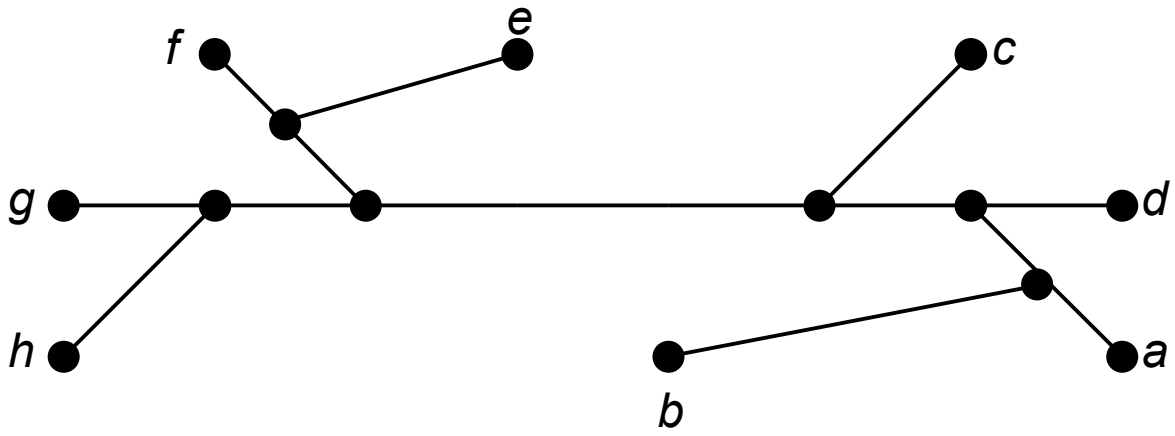
An agreement forest for two trees T,T' in $\boldsymbol{T}_n$ is a collection $\{T_0,\ldots,T_k\}$ of binary phylogenetic trees such that

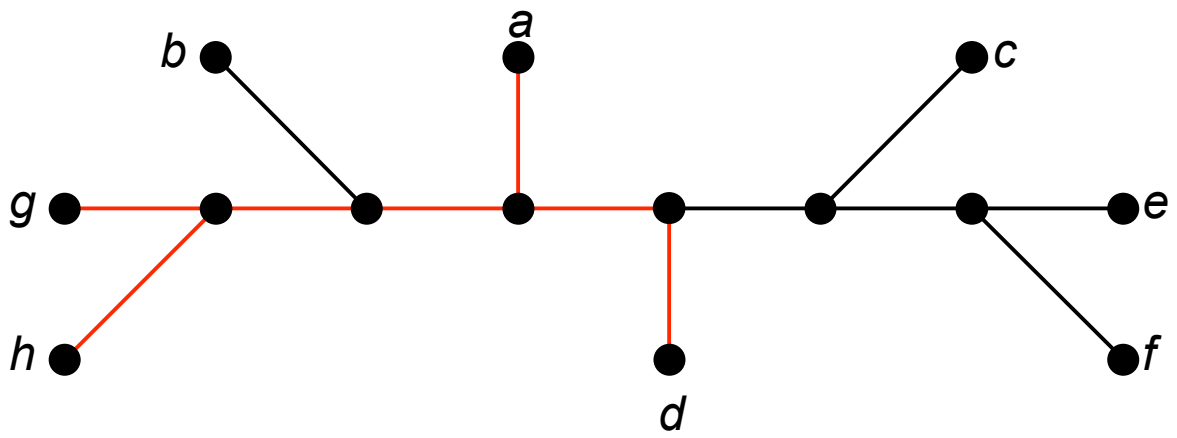
(i) the taxa sets of $T_0,\ldots,T_k$ form a partition of $\{1,\ldots,n\}$.

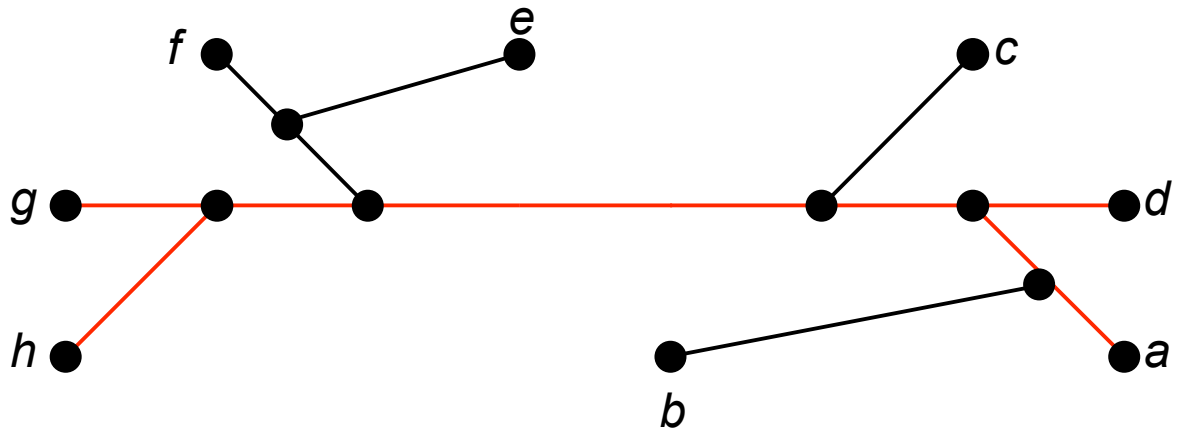
(ii) T_i is a restriction of T and T' for all i.

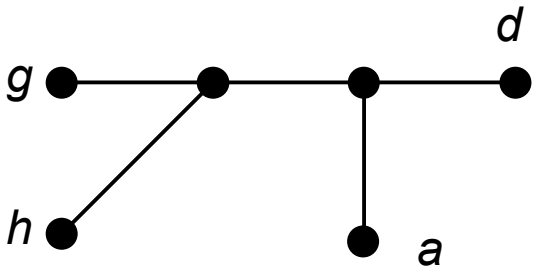
(iii) The smallest subtrees containing $L(T_0),\ldots,$ $L(T_k)$ of T resp. T' are vertex-disjoint.

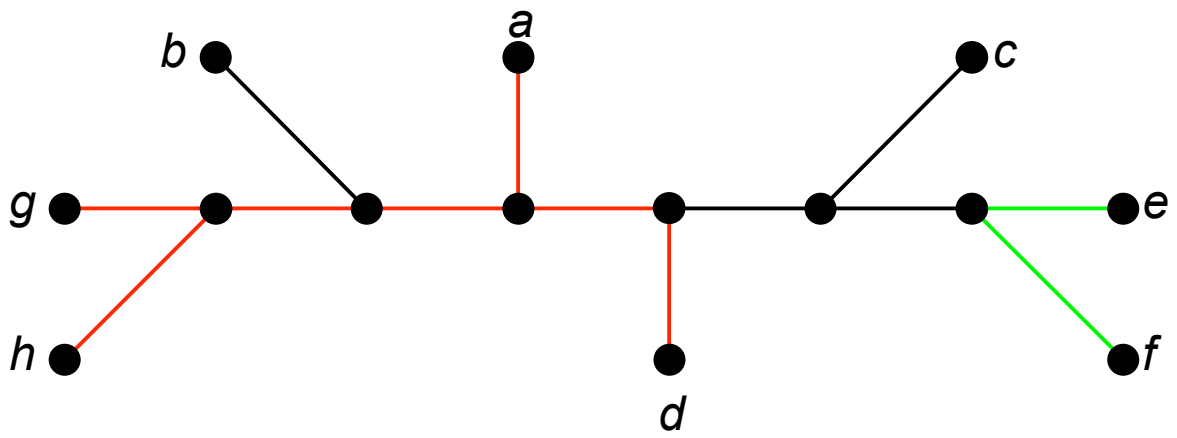

a
b
c
g
e
h
f
d

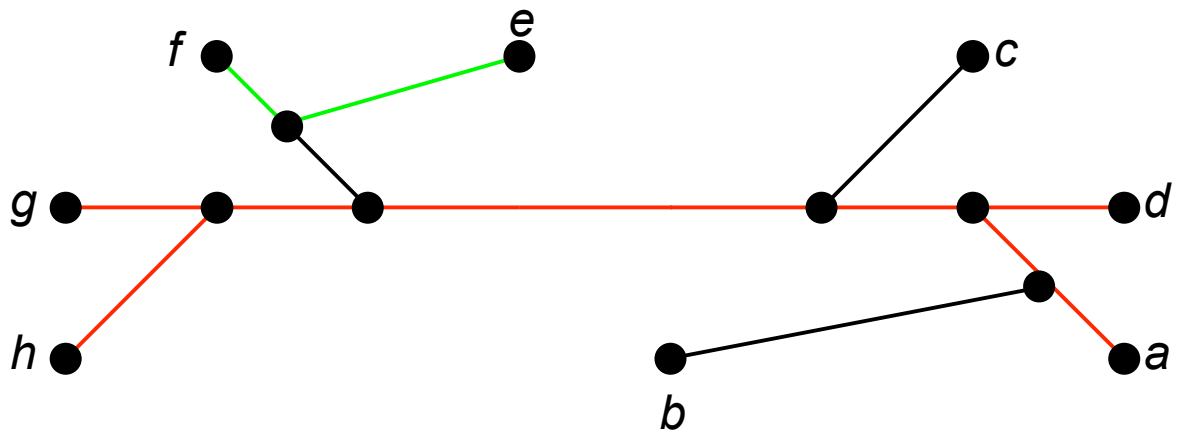

f
e
c
g
d
h
a
b

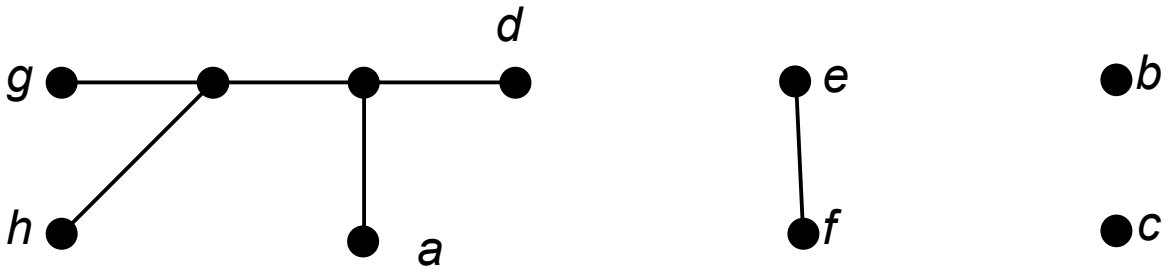

a
b
c
g
e
h
f
d


e
f
c
g
d
h
a
b


d
g
h
a


a
b
c
g
e
h
f
d


f
e
c
g
d
h
a
b


d
g
h
a

e
f

b
c

Maximum agreement forests

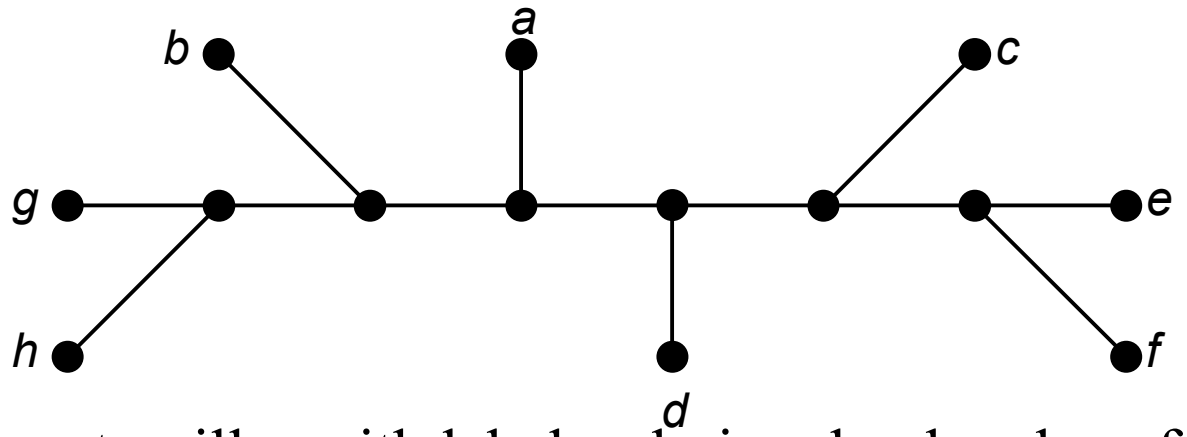
- An agreement forest for T,T' is a *maximum agreement forest* if the number of trees is minimal.

Lemma 1 (Allen, Steel, 2001): If $\{T_0,\ldots,T_k\}$ is a maximum agreement forest for T,T', then $d_{\mathrm{TBR}}(T,T')=k$.

Lemma 2 : If $\{T_0,\ldots,T_k\}$ is an agreement forest for T,T' such that every tree contains at most 2 taxa, then $d_{SPR}(T,T')\leq k$.

Caterpillars

A *caterpillar* is a binary phylogenetic tree where the interior vertices form a path. A *label ordering* is a permutation of the taxa set such that two consecutive elements are adjacent to the same interior vertex or to two adjacent interior vertices.



A caterpillar with label ordering h,g,b,a,d,c,e,f.

The lower bound

Lemma 3: Let k, l be positive integers such that $2 \leq k \leq l$, and let T , $T' \in \boldsymbol{T}_{kl}$ be caterpillars such that T has the label ordering $[1, \ldots, kl]$ and T' has the label ordering $[1, k+1, \ldots, k(l-1) + 1, 2, k + 2, \ldots, k(l-1) + 2, \ldots, k, k + k, \ldots, k(l-1)+ k]$. Then $d_{\mathrm{TBR}}(T,T')=(k-1)(l-1)$.

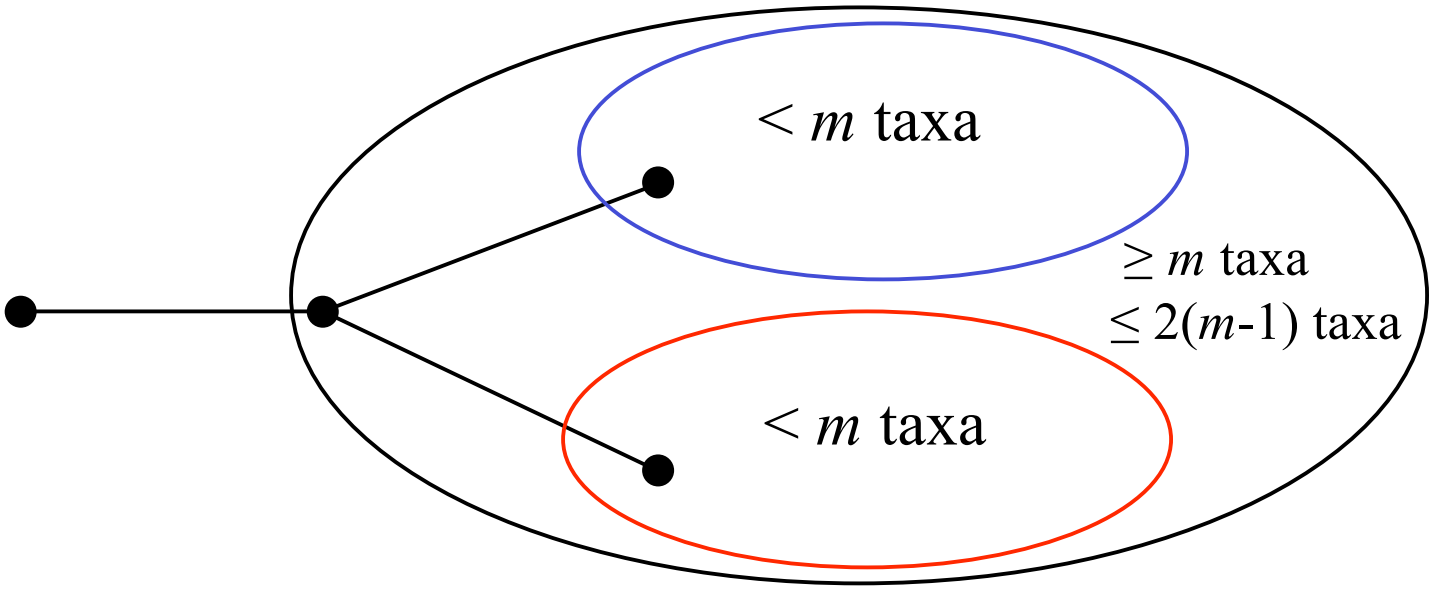
To obtain the lower bound we choose $k \approx l$.

Chopping trees

Lemma 4: Let $k \geq 0$ and $l, m, n > 1$ be integers such that $n \geq 2k(m-1) + l$, and let $T \in \mathcal{T}_n$ Then there is a collection $T_0,\ldots,T_k$ of vertex-disjoint subtrees of T such that $|L(T_0)| \geq l$ and $|\mathrm{L}(T_i)| \geq \mathrm{m}$ for all $\mathrm{i} \in \{1,...,\mathrm{k}\}$.

Chopping trees

Lemma 4: Let $k \geq 0$ and $l, m, n > 1$ be integers such that $n \geq 2k(m\text{-}1) + l$, and let $T \in \boldsymbol{T}_n$ Then there is a collection $T_0,\ldots,T_k$ of vertex-disjoint subtrees of T such that $|L(T_0)| \geq l$ and $|\mathrm{L}(T_i)| \geq \mathrm{m}$ for all $\mathrm{i} \in \{1,\ldots,\mathrm{k}\}$.

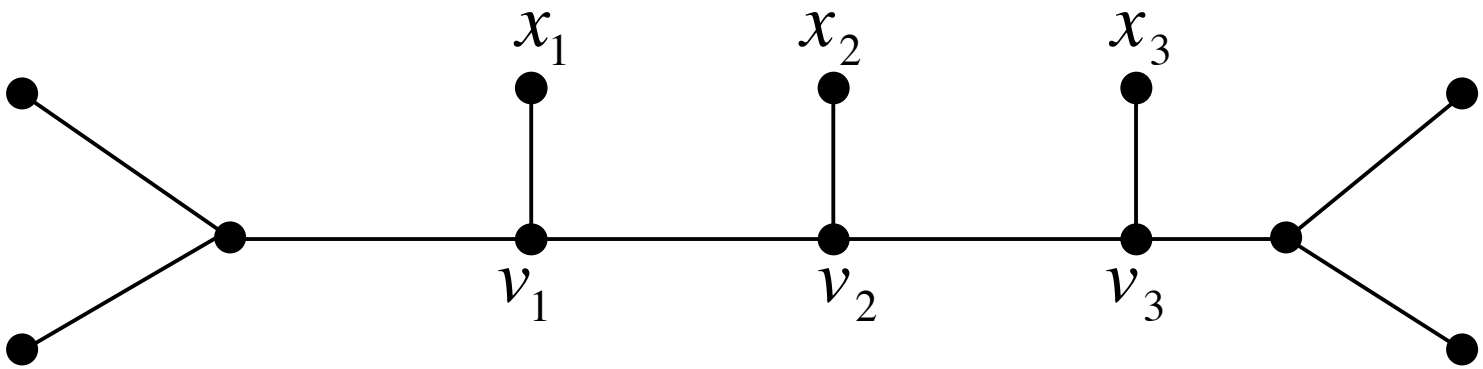


The upper bound

- Given T, T' in $\boldsymbol{T}_n$,
- We chop T into about $\sqrt{n}$ trees with about $\sqrt{n}$ taxa.
- Then we chop smallest possible trees from T' such that the chopped tree has at least taxa with one of the subtrees of T (which has not yet been used) in common.
- We get an agreement forest with about $\sqrt{n}$ trees with 2 taxa.
- Applying Lemma 2 yields the upper bound.

Chains

A chain of length l in a phylogenetic tree is a path $v_1,...,v_l$ of l interior vertices such that every vertex v is adjacent to a leaf x_i (i=1,…,l).



The Chain Reduction Conjecture

Conjecture (Allen, Steel 2001): If two binary phylogenetic X-trees T and T' both contain the same chain of length $l \geq 4$, then the SPR distance does not change if the chain is replaced by identical chains of length 3 in both trees (correctly oriented).

Consequences

- The corresponding result holds for TBR (easy to prove using maximum agreement forests)
- The conjecture implies fixed-parameter tractability of computing the SPR distance between two given trees.
- This has been shown using a different approach.

More reasons to solve it

The chain reduction is already implemented in a program to compute (or estimate) the SPR distance (Hickey et al. 2008).

They also gave statistical evidence by testing 20000 pairs of trees.

More reasons to solve it

The chain reduction conjecture is one of Mike Steel's 100 NZ$ problems.

It even became a Penny ante and solving it yields a bottle of single malt.

Induced SPR sequences

Every sequence S of SPR operations between two X-trees T and T' defines a sequence between the restrictions of T and T' to a subset X' of X.

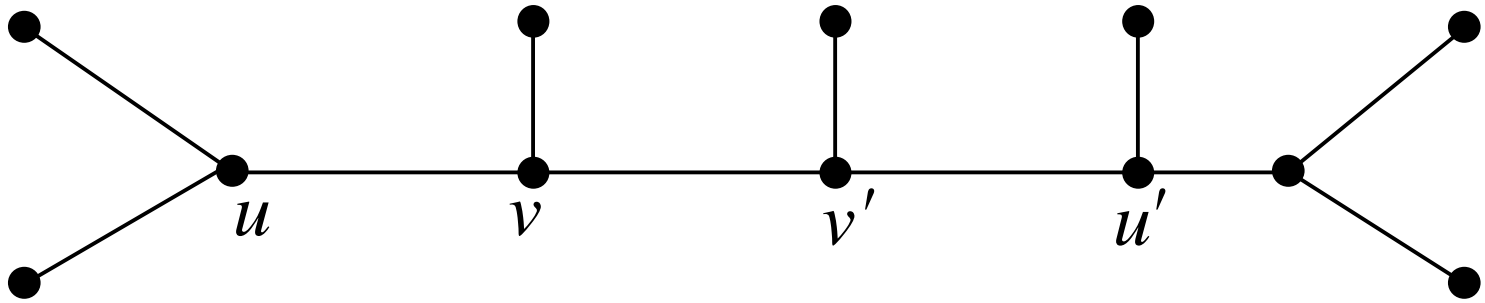
If two trees are identical, then the operation is removed from the sequence. Hence,

$$d_{\mathrm{SPR}}(T_{|X'}, T'_{|X'}) \leq d_{\mathrm{SPR}}(T, T')$$

A reformulation

- We fix two X-trees T and T' and edges uv and $u'v'$, respectively.
- We denote the trees that we get by subdividing the edge uv resp. $u'v'$ by a chain of length i with taxa $x_1,...,x_l$ with increasing indices from u to v (resp. u' to v') by T_i resp. T_i'.
- We define $d_i = d_{\mathrm{SPR}}(T_i, T_i')$.
- **Conjecture:** $d_i = d_3$ for every integer $i \geq 3$.

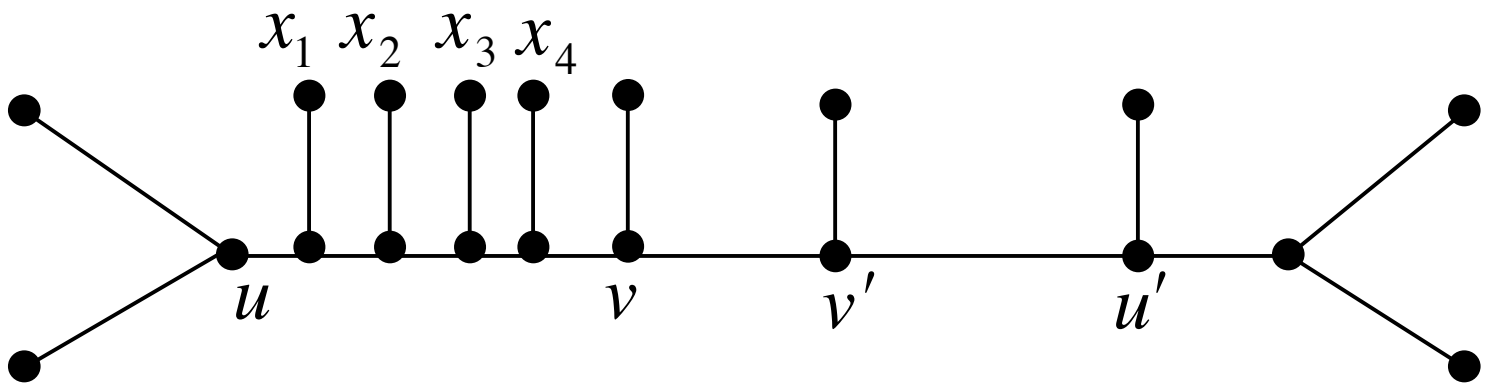
An Example



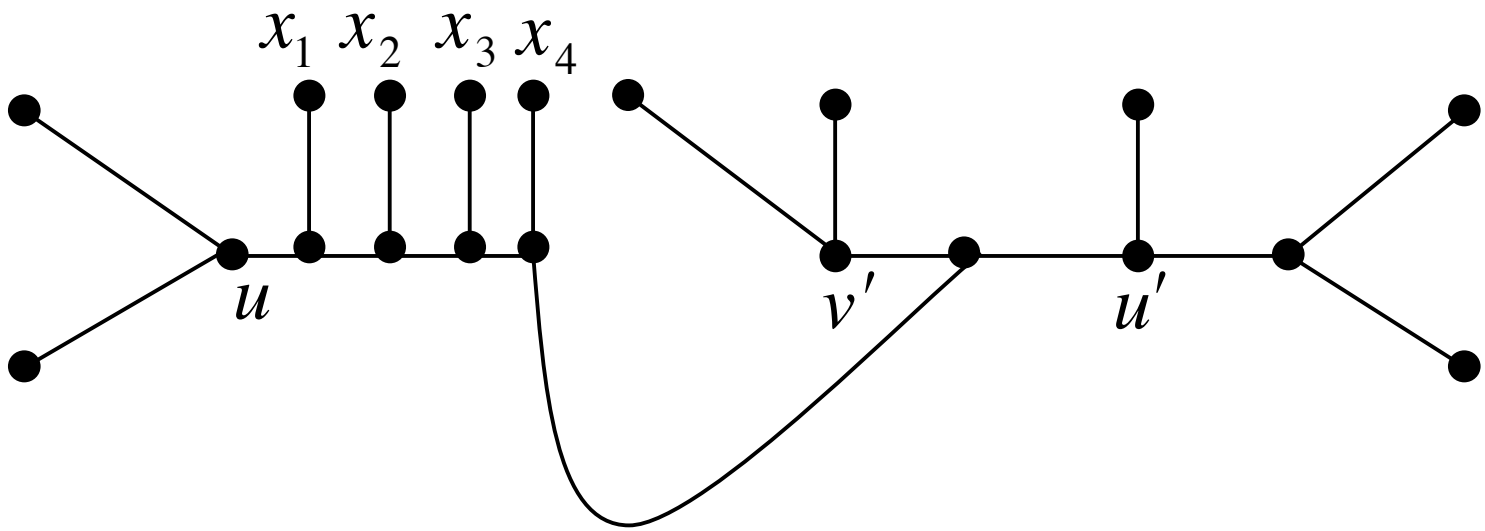
Let $T=T'$ and u,v,u',v' as above. We have

$d_0 = 0$, $d_1 = 1$, $d_2 = 2$, $d_3 = 3$, and $d_i = 3$ for $i > 3$.

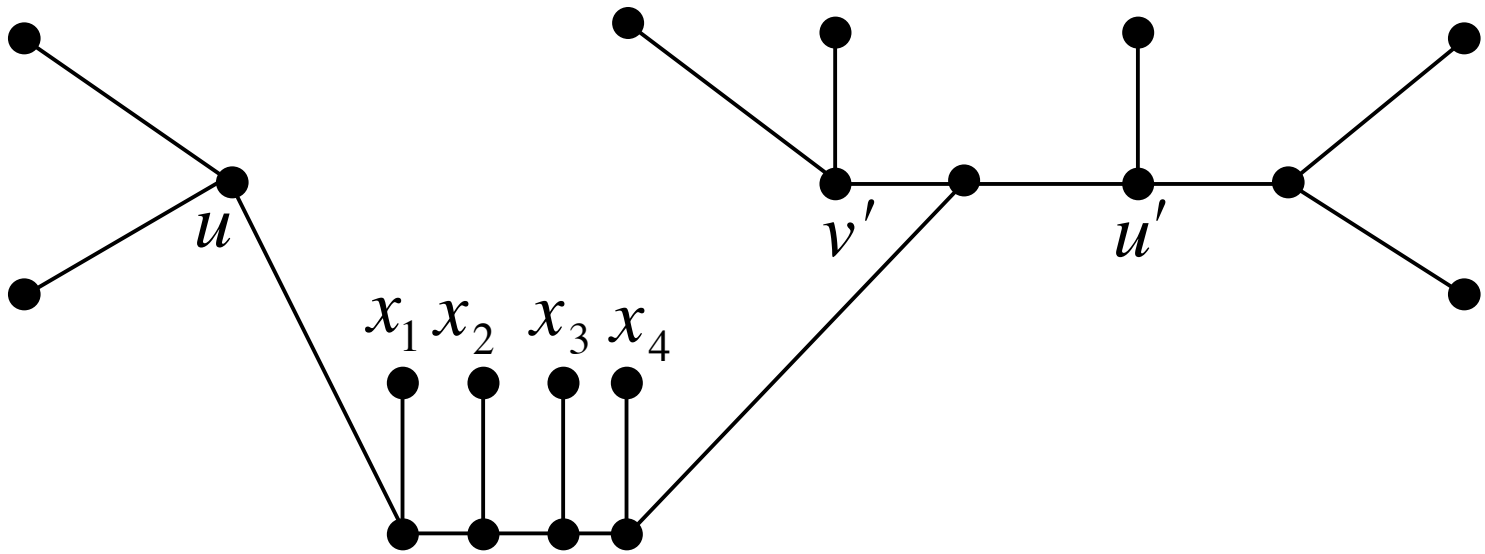
An Example



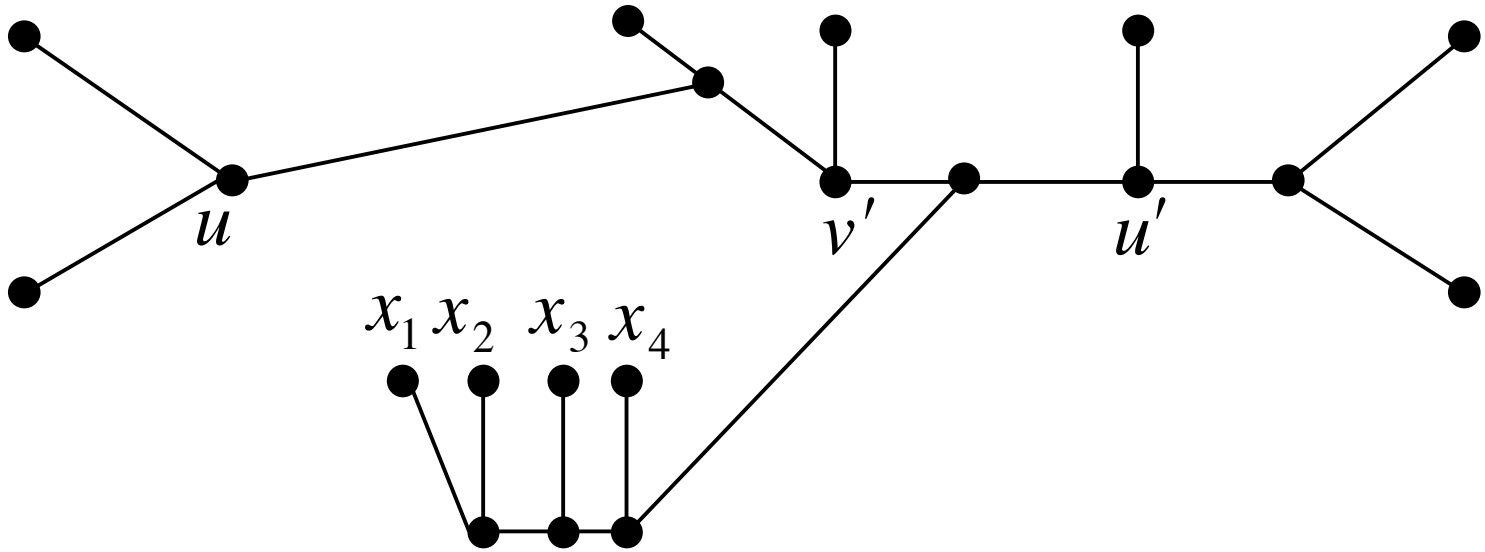
An Example



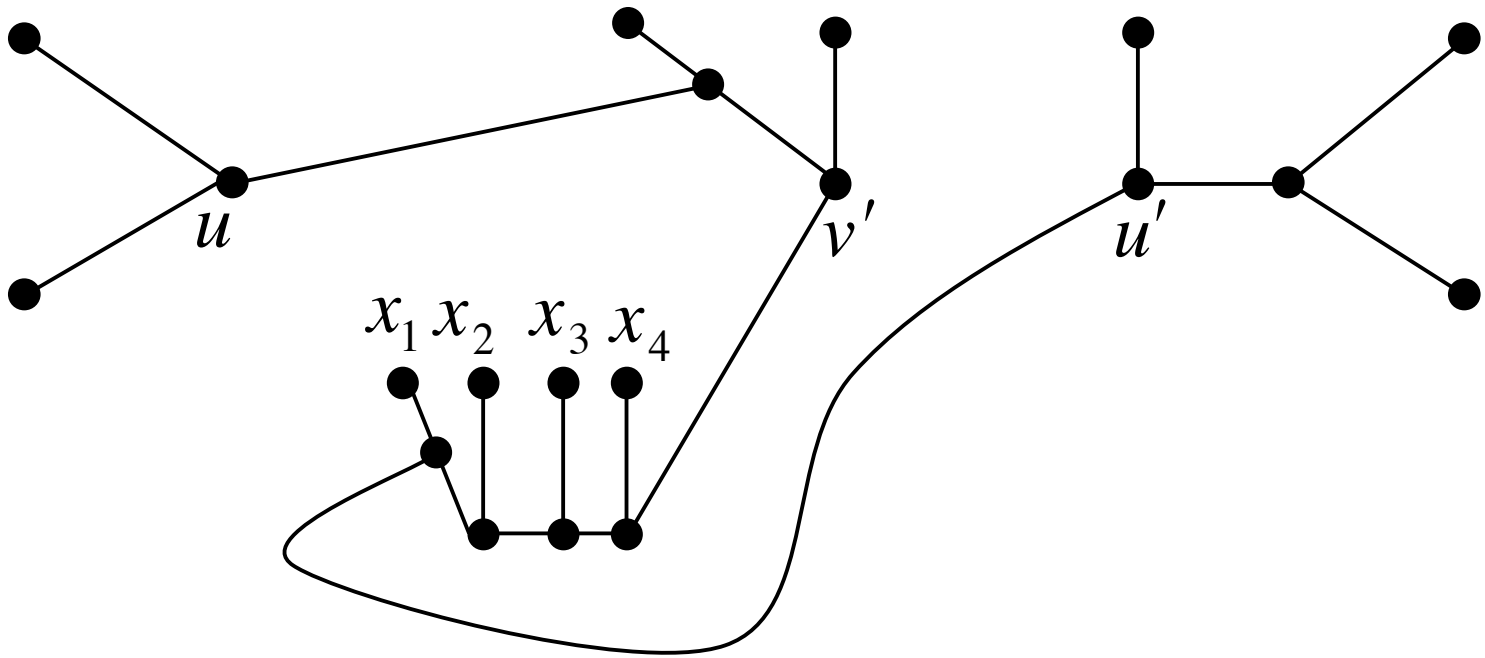
An Example



An Example



An Example



An easy lemma

Lemma 5: $d_i \leq d_0+3$ for all i.

Statement: If $d_i = d_{i+1}$ for some $i \geq 1$, then $d_j = d_i$ for every $j \geq i$.

An easy lemma

Lemma 5: $d_i \leq d_0+3$ for all i.

~~**Statement:** If $d_i = d_{i+1}$ for some $i \geq 1$, then $d_j = d_i$ for every $j \geq i$.~~

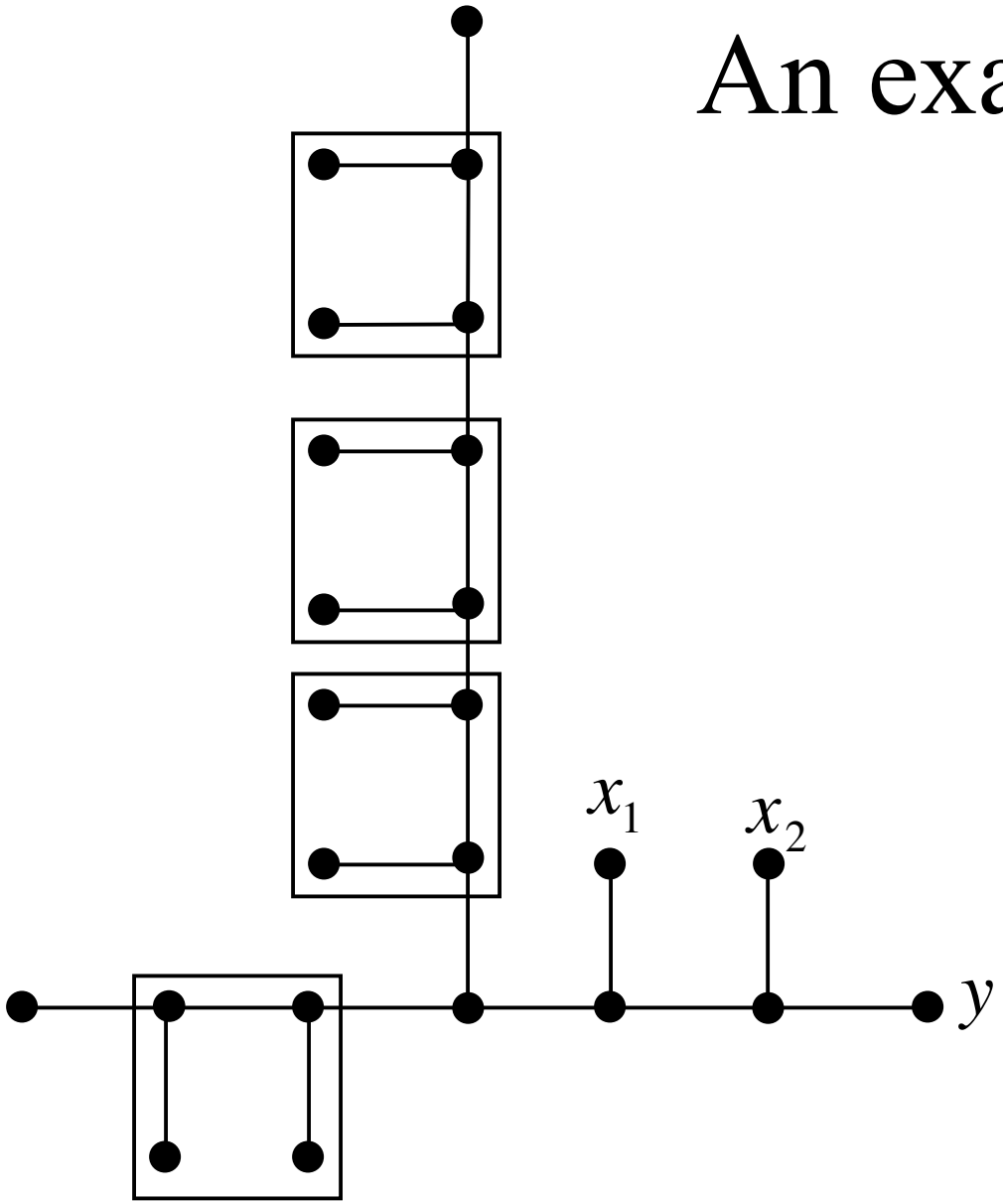
Very long chains

- **Theorem (Bonet, St. John 08):** There is a linearly bounded function f:**N**$\rightarrow$ **N** such that $d_0 \leq d$ implies $d_i = d_{f(d)}$ for every integer $i \geq f(d)$.
- Using their ideas we can show that for two trees T,T' in $\boldsymbol{T}_n$ there is a shortest SPR sequence such that all edges in a chain of length $f(d)$ are never altered (removed or subdivided).

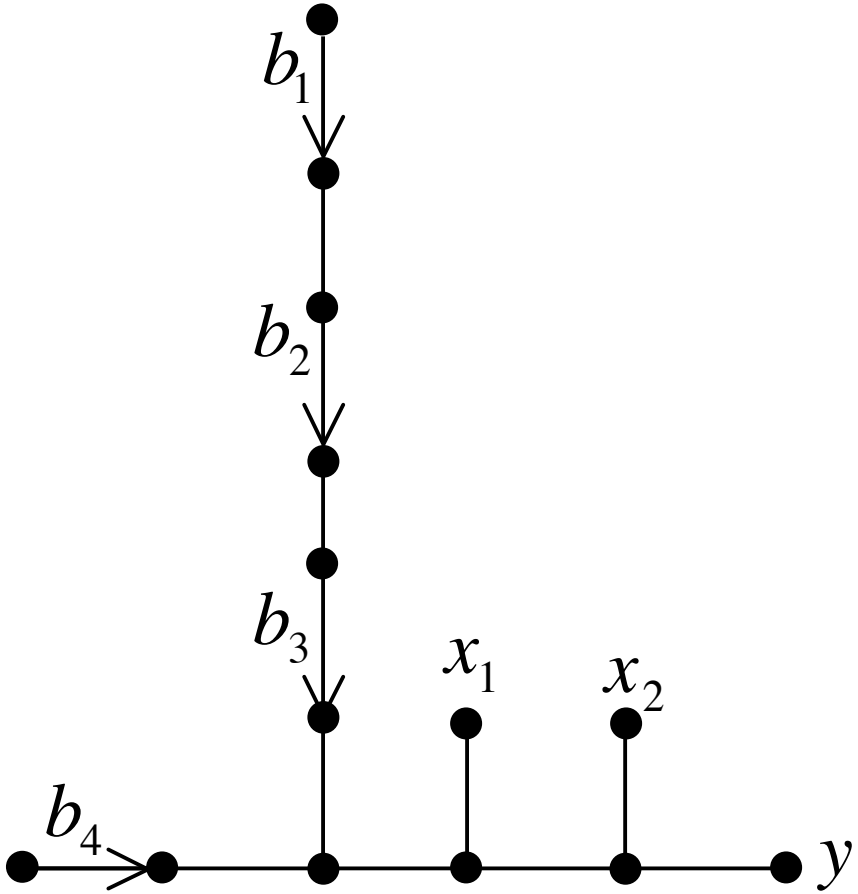
Blocks

- In order to find and verify a counterexample, we want to exclude most possible moves (otherwise exhaustive search is not feasible).
- We do so by inserting sufficiently long chains and replacing them by *blocks,* that is chains of length 2 that must not be altered.

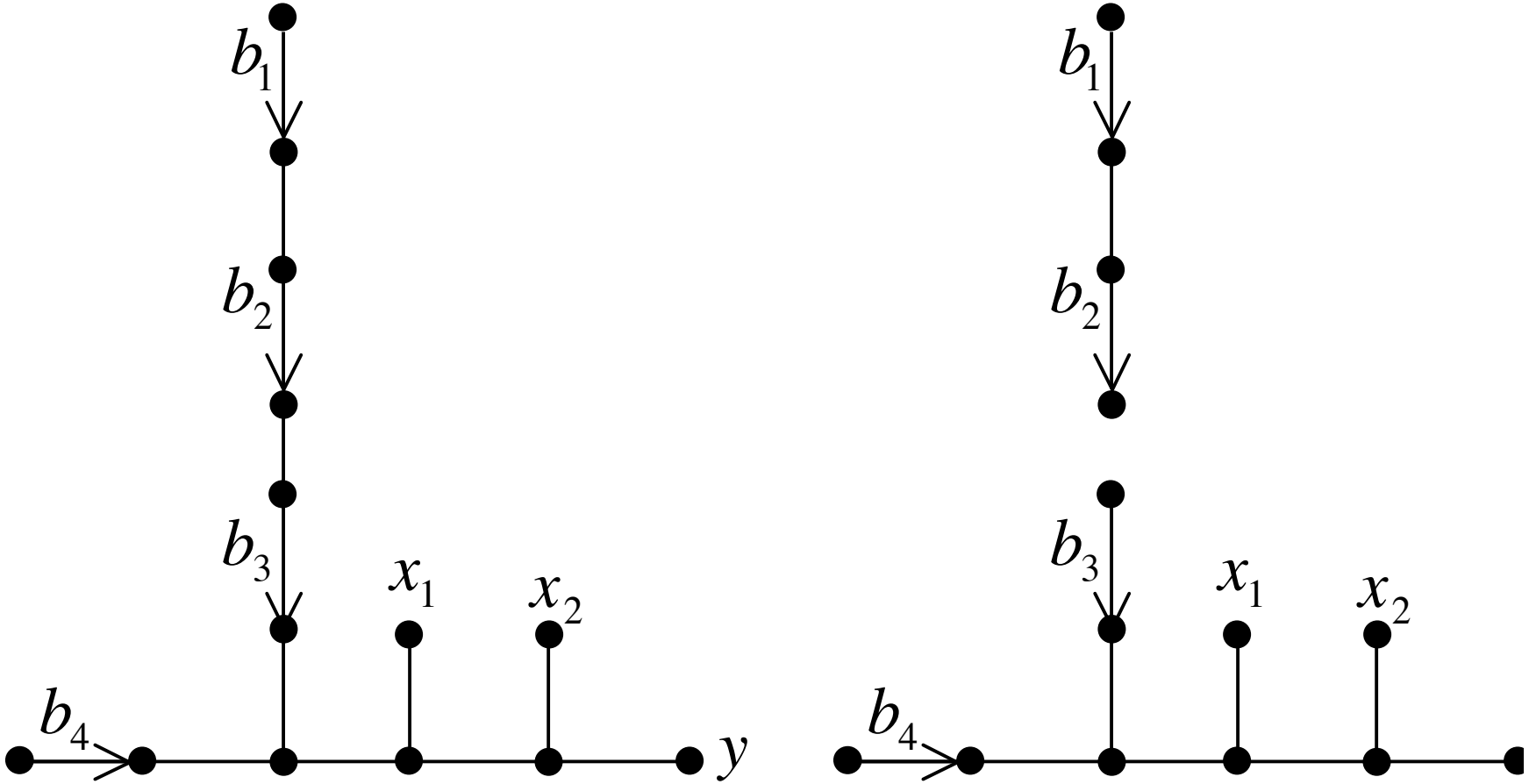
An example



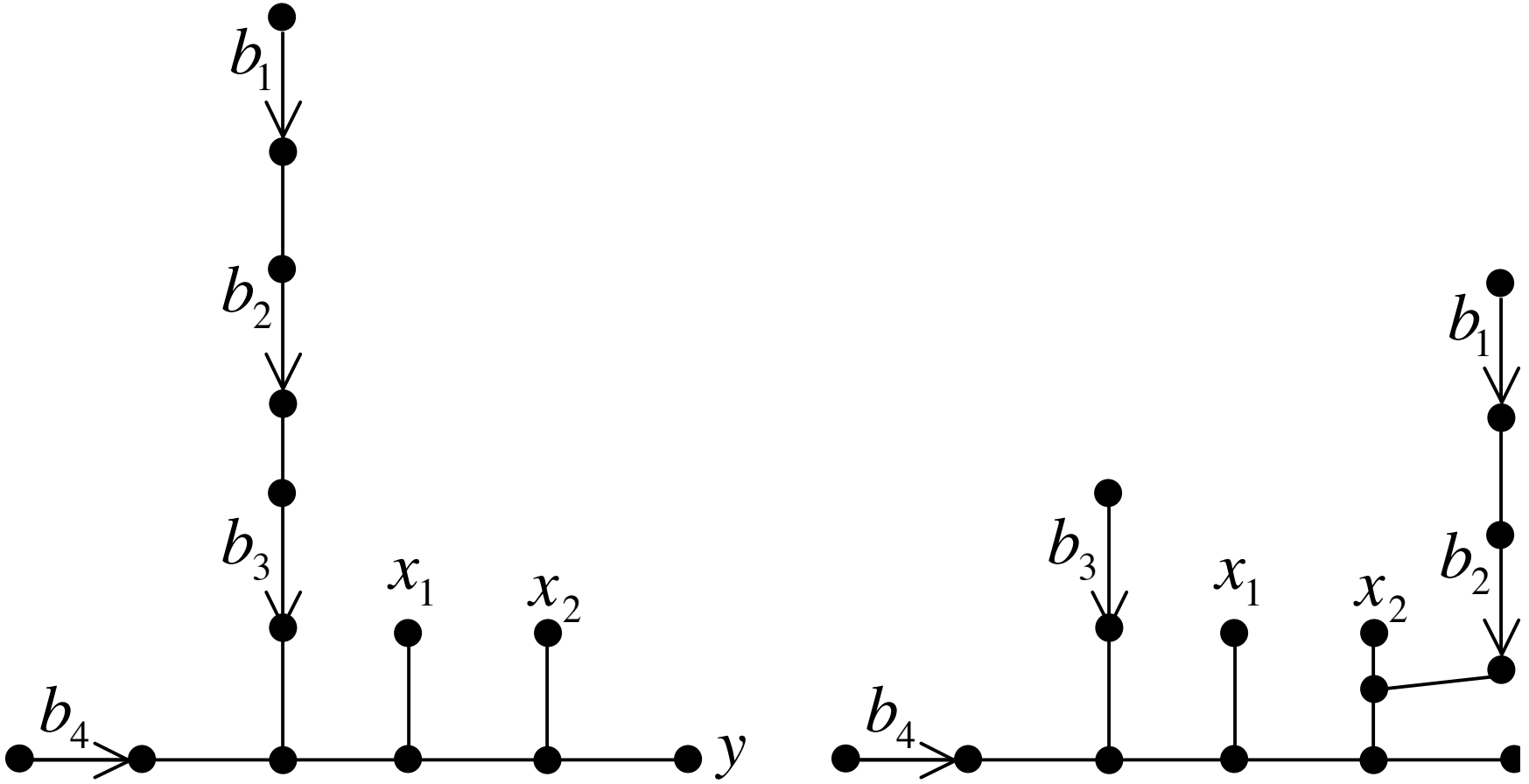
An example



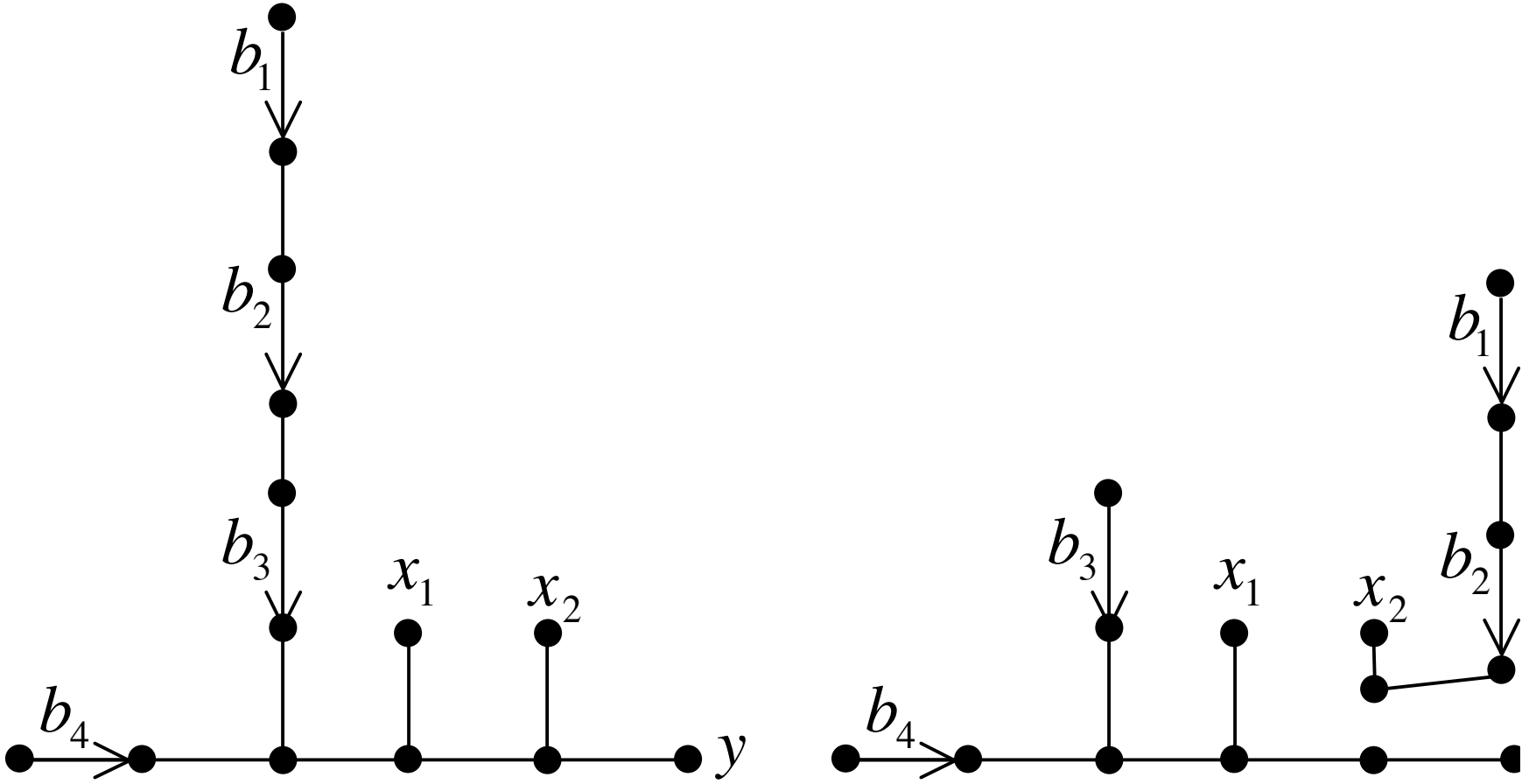
An example



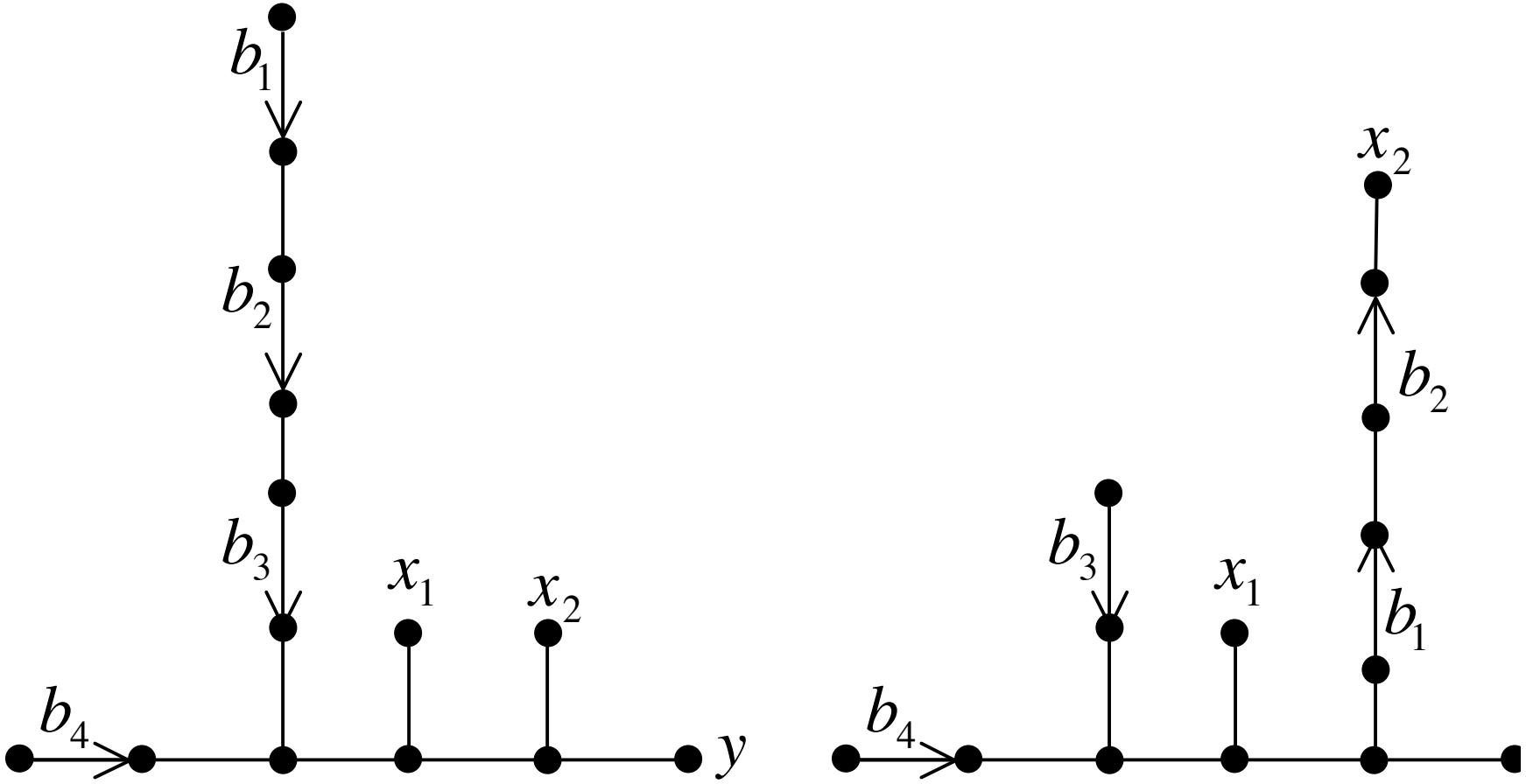
An example



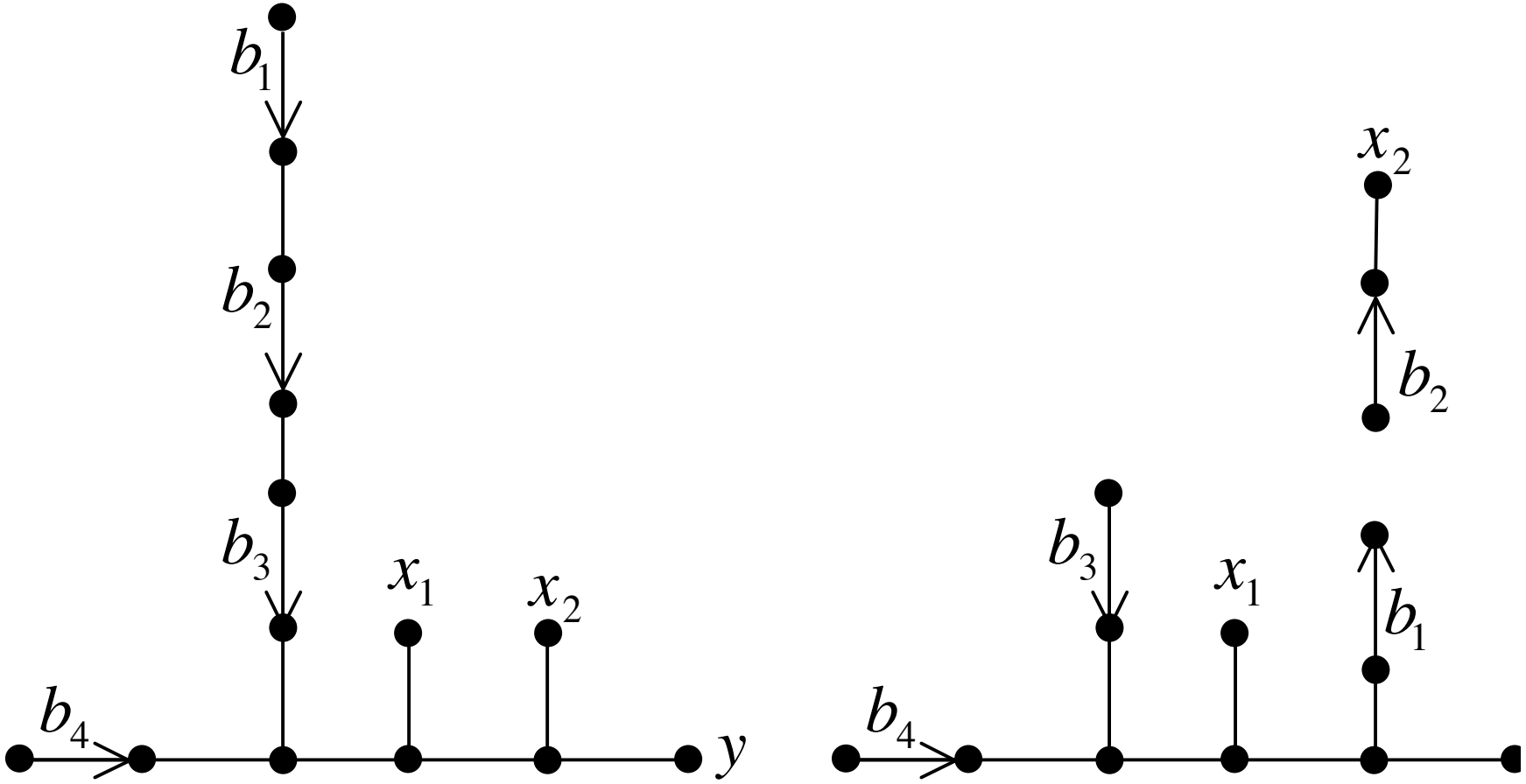
An example



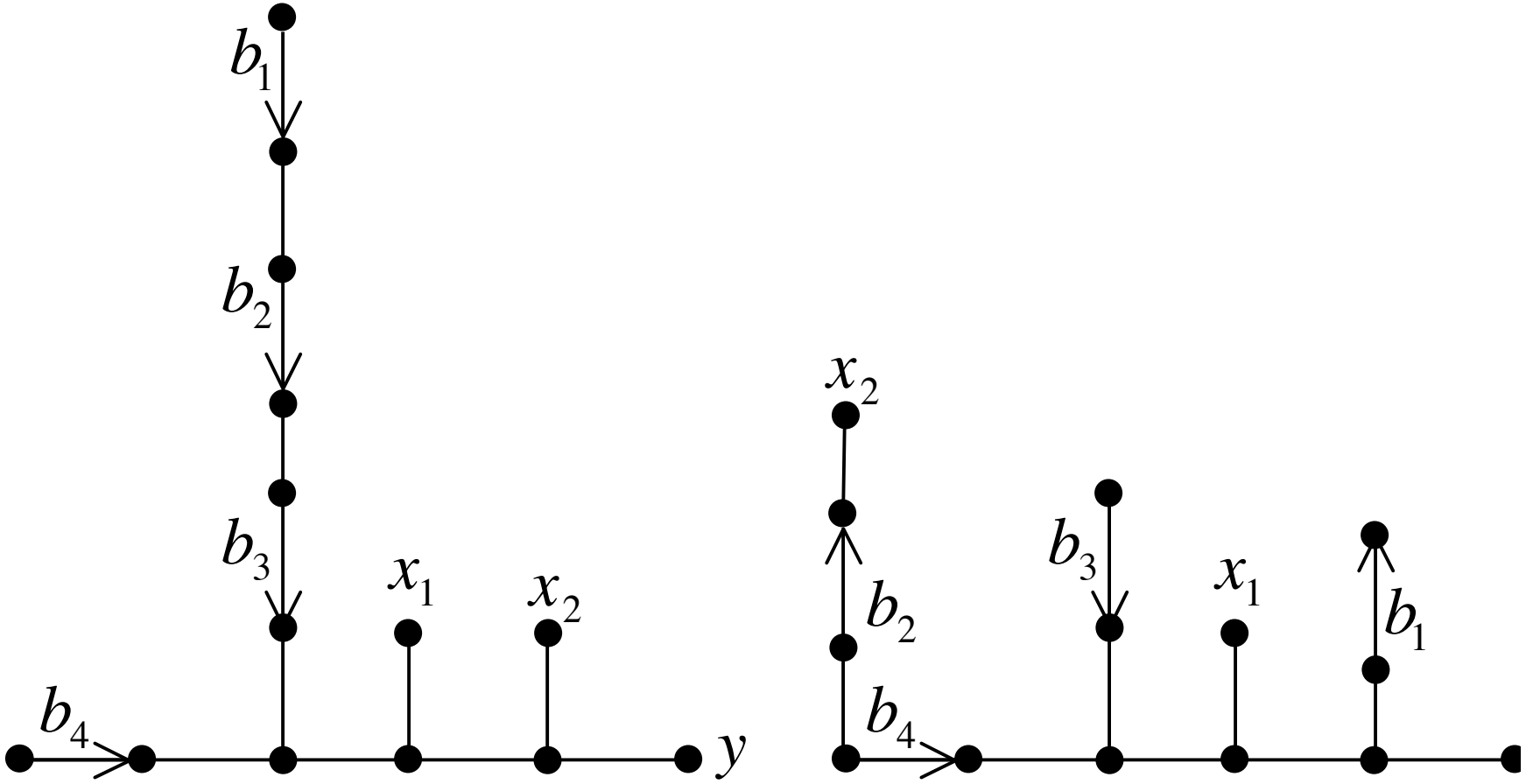
An example



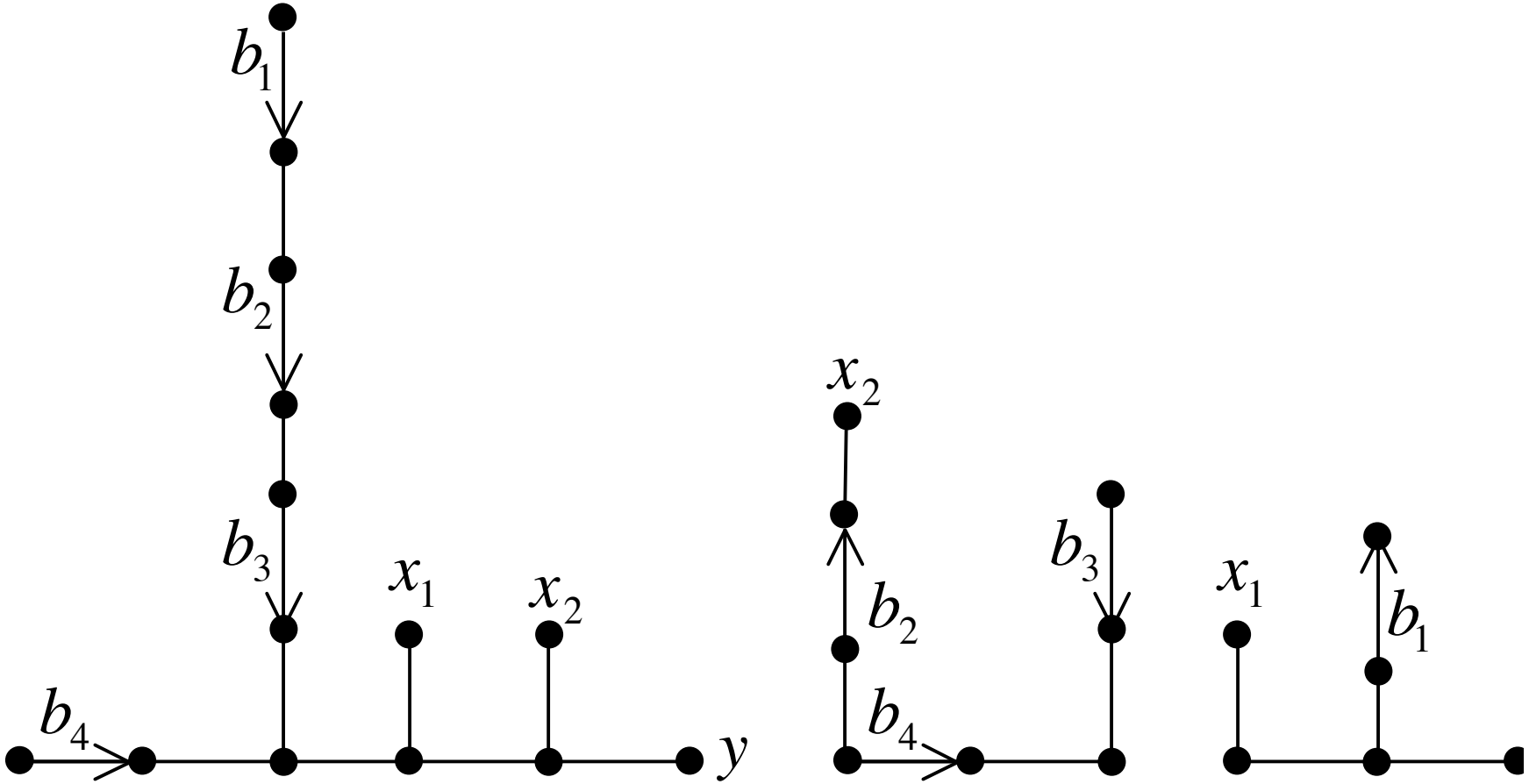
An example



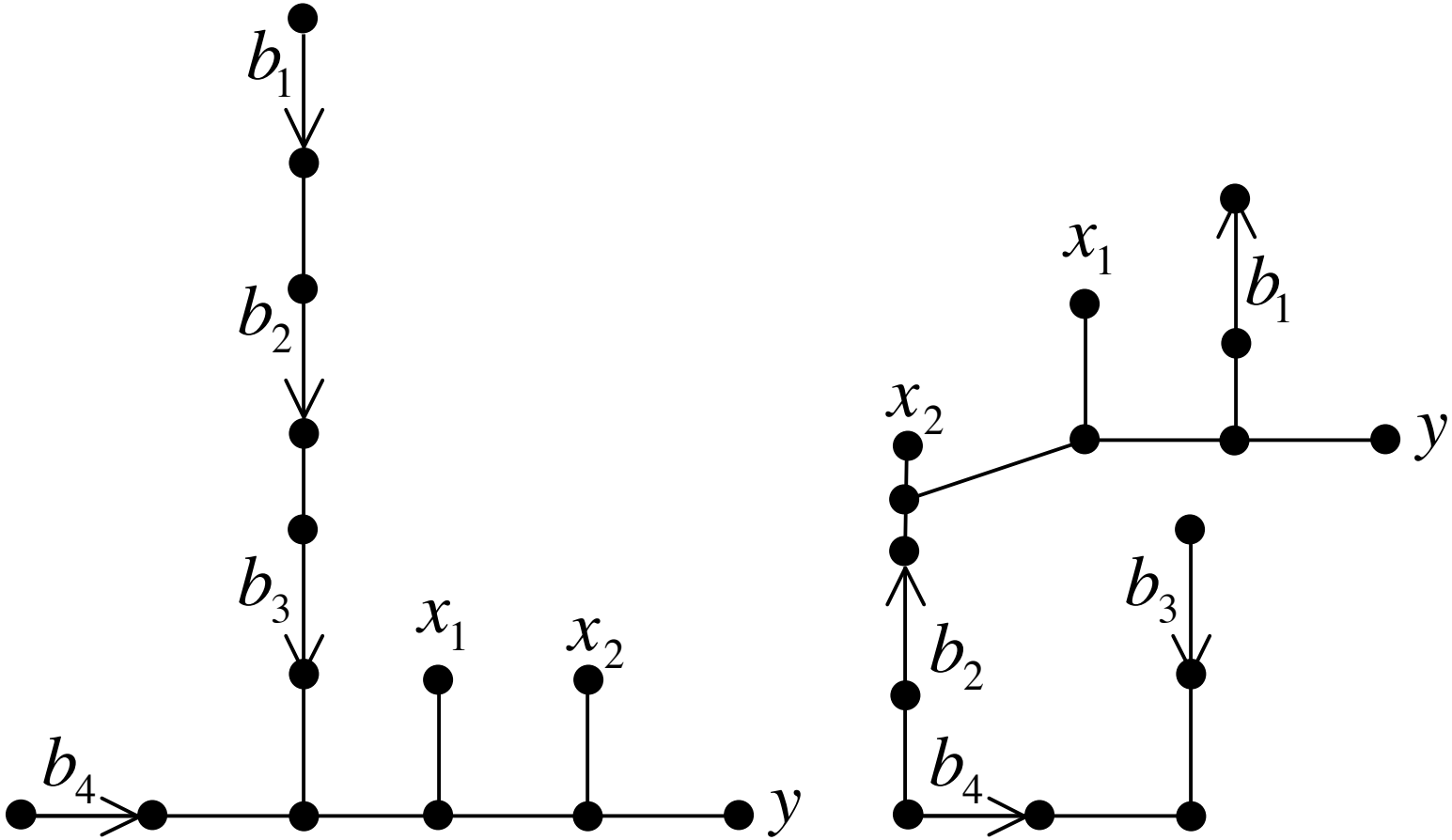
An example



An example



An example



A lower bound

- Given T, T' in $\mathcal{T}_n$ with identical blocks. Removing all block edges defines 2 partitions P, P' of the taxa (one for each tree).
- Let k be the smallest number of parts in a partition that refines both, P and P'.
- Then $d_{\mathrm{SPR}}(T_i, T_i') \geq k - \max\{P, P'\}$.

Not there, yet

- We have implemented a program that can compute the SPR distance for two trees with blocks using the lower bound above.
- We have examples that gave us a lot of insights into a problem.
- However, the conjecture is still open.

Acknowledgment

- The part on the chain reduction conjecture is joint work with Jun Li.

Acknowledgment

- The part on the chain reduction conjecture is joint work with Jun Li.

Thank you!