Reconstructing the parameters of a network from its tree-average distances

Stephen J. Willson

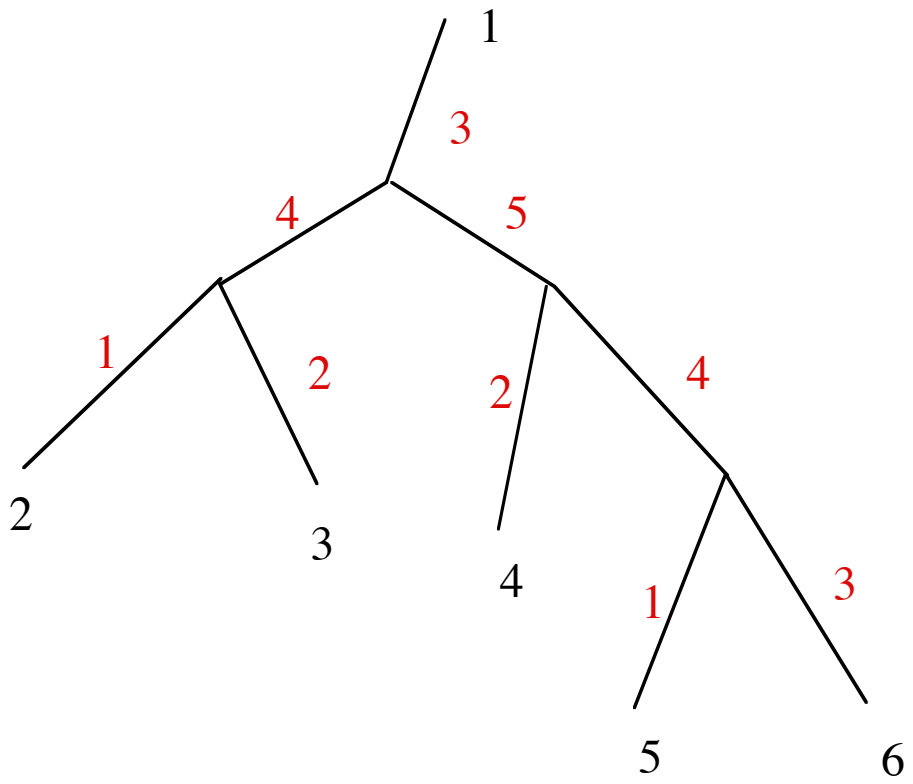
Abstract. A phylogenetic network is a rooted acyclic directed graph with labeled leaves which correspond to extant species. The network depicts the course of evolutionary history as species mutate and hybridize. Often each arc is weighted by a nonnegative real number measuring the amount of genetic change along the arc. A "tree-average distance" is defined which tells the average of the distances between the leaves in each displayed tree using these weights. Given a normal network N with certain restrictions, we show how the weights may be reconstructed from the tree-average distances. With additional assumptions we also indicate how the network N itself may be reconstructed.

Reconstructing the parameters of a network from its tree-average distances

Stephen J. Willson

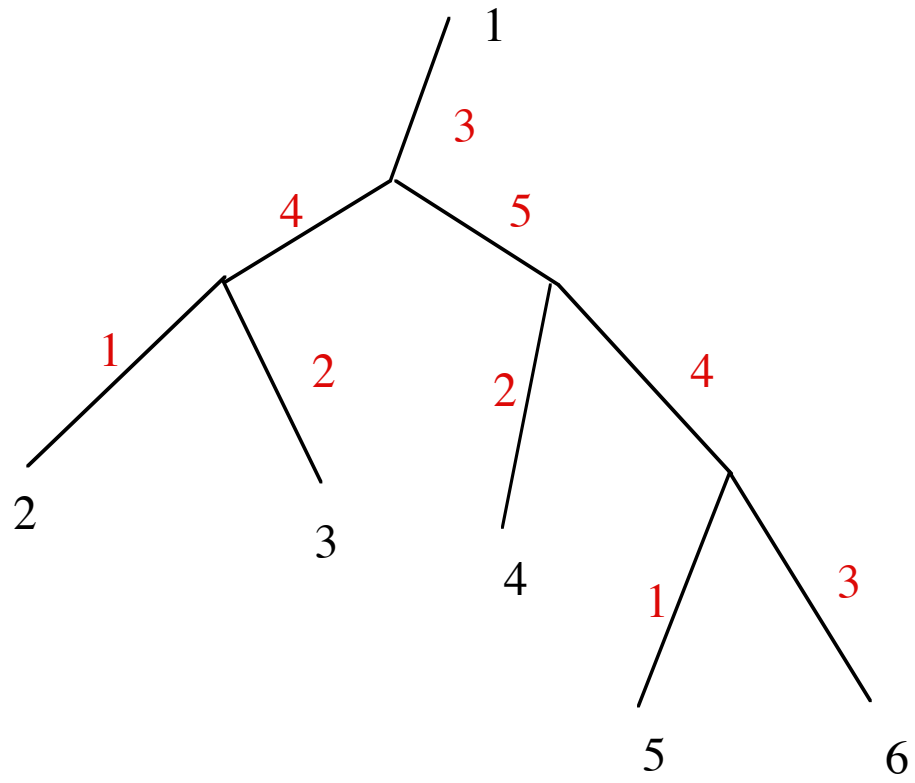
Phylogenetics: New data, new phylogenetic challenges
Isaac Newton Institute
Cambridge UK
21 June 2011

Often the arcs of a phylogenetic tree have a nonnegative **weight** that measures the amount of genetic change along the arc.



A distance function d is **additive** if d(x,y) is the sum of the weights on the path between x and y.

$d(x,y) = \sum$ [w(e): arc e lies on the path between x and y]

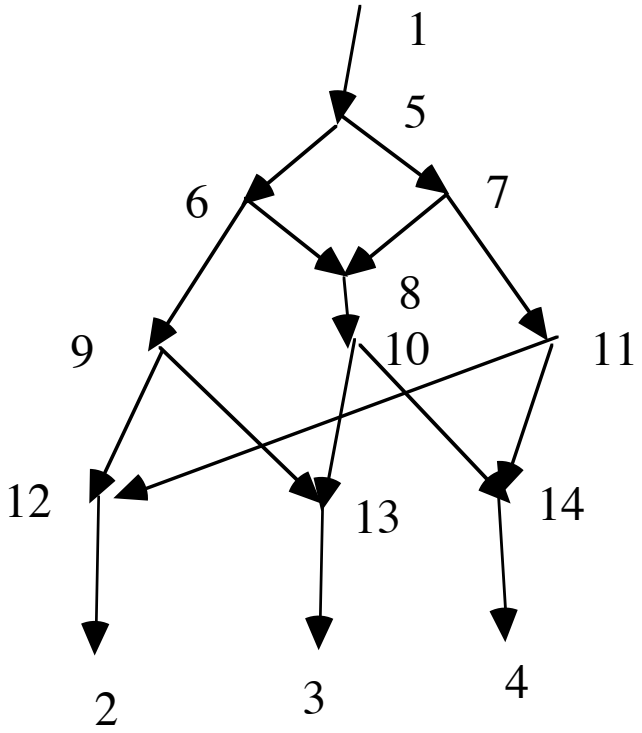


d(2,4) = 1+4+5+2 = 12.

Theorem. Suppose no vertex has total degree 2. Suppose d is an additive distance on the tree T. Then the tree T is uniquely determined and the weights are uniquely determined. Explicit formulas can be given for the weights once T is known.

Distance methods are very fast and efficient. Neighbor-joining (Saitou and Nei 1987) is robust in the presence of errors in the distances.

What is a generalization to rooted directed acyclic networks?

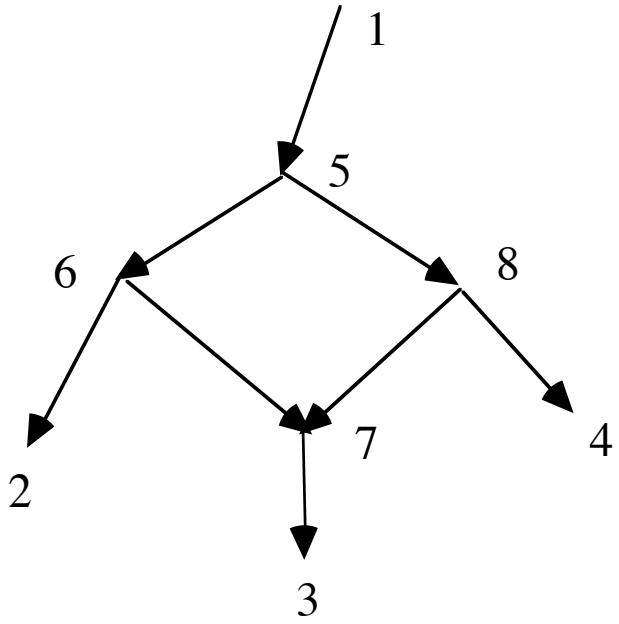


A network can include also hybridization events.

Assume that the distances are known between members of X = {1,2,3,4}.
There are C(4,2) = 6 distances but 17 arcs. The lengths of the 17 arcs cannot be determined uniquely from the distances between members of X.

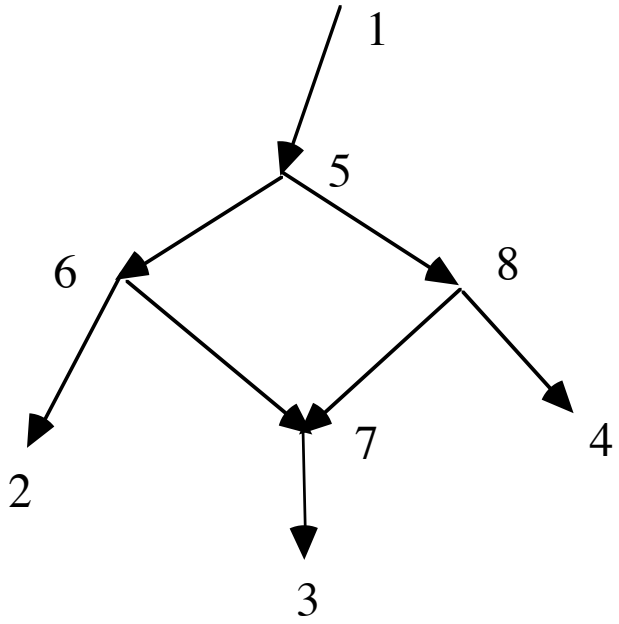
If there are more arbitrary parameters than distances between members of X, the parameters cannot be uniquely determined.

We need to restrict the network.



A network with X = {1,2,3,4}. There are C(4,2) = 6 distances. There are 8 arcs.

A vertex is **normal** or **tree-child** if it has indegree 1.
A vertex is **hybrid** if it has indegree greater than 1.
An arc (a,b) is **normal** if b is normal.



A network with X = {1,2,3,4}. There are C(4,2) = 6 distances. There are 8 arcs.

Assume: If (a,b) is an arc with b hybrid, then w(a,b) = 0.

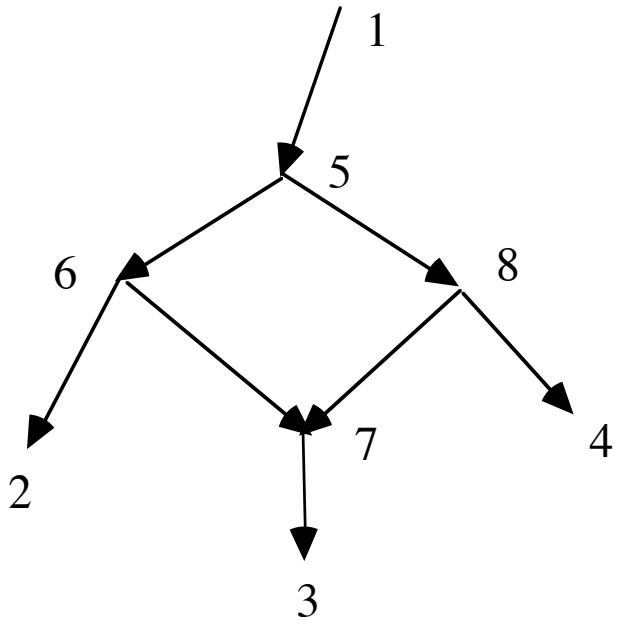
Now in the example there are 6 distances and 6 parameters--the weights of the 6 normal arcs that don't enter a hybrid vertex.

Assume for the remainder of this talk:

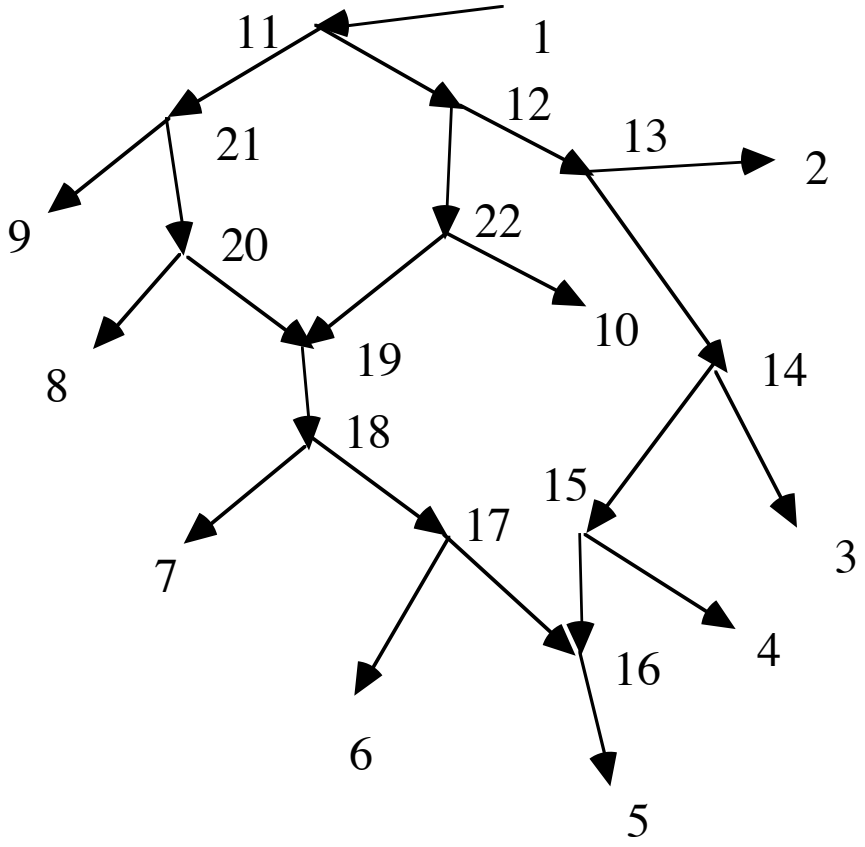
(1) Each arc (a,b) has a weight w(a,b).

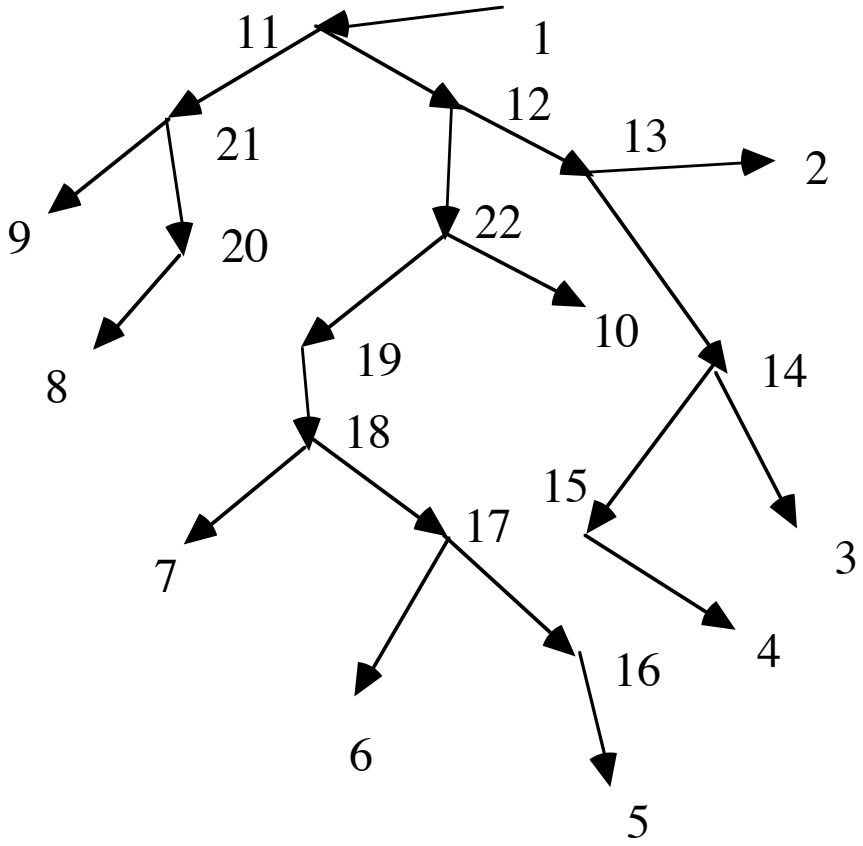
(2) If (a,b) is an arc with b hybrid, then w(a,b) = 0.

(3) Each hybrid vertex has exactly one child, and this is a tree-child.

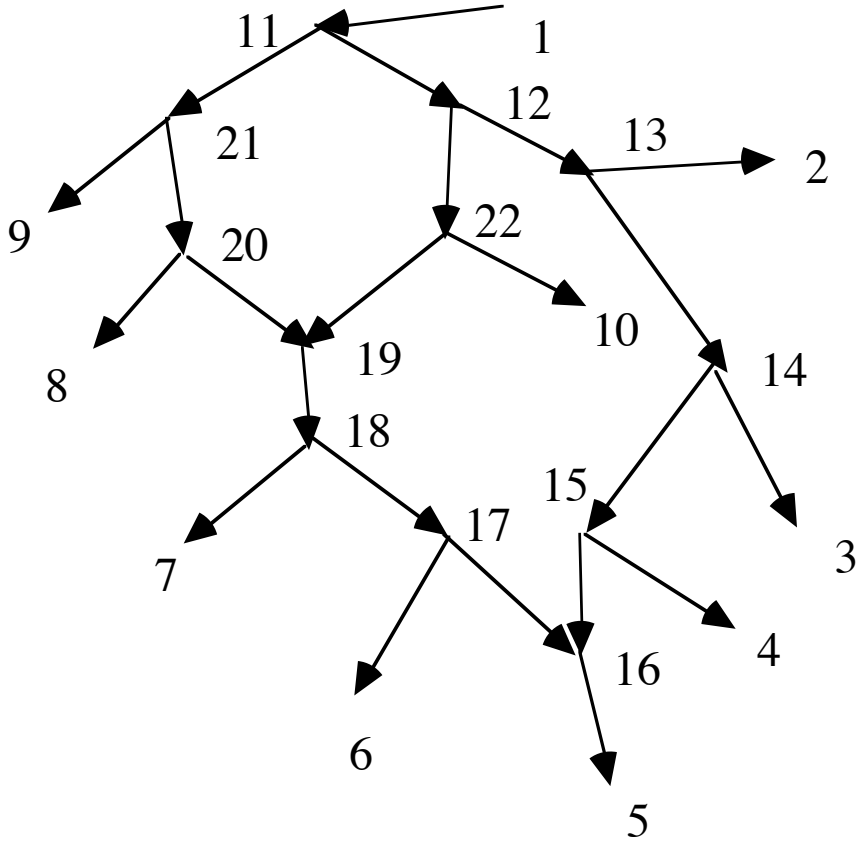


Biological assumption: Various genes follow different trees displayed by the network.



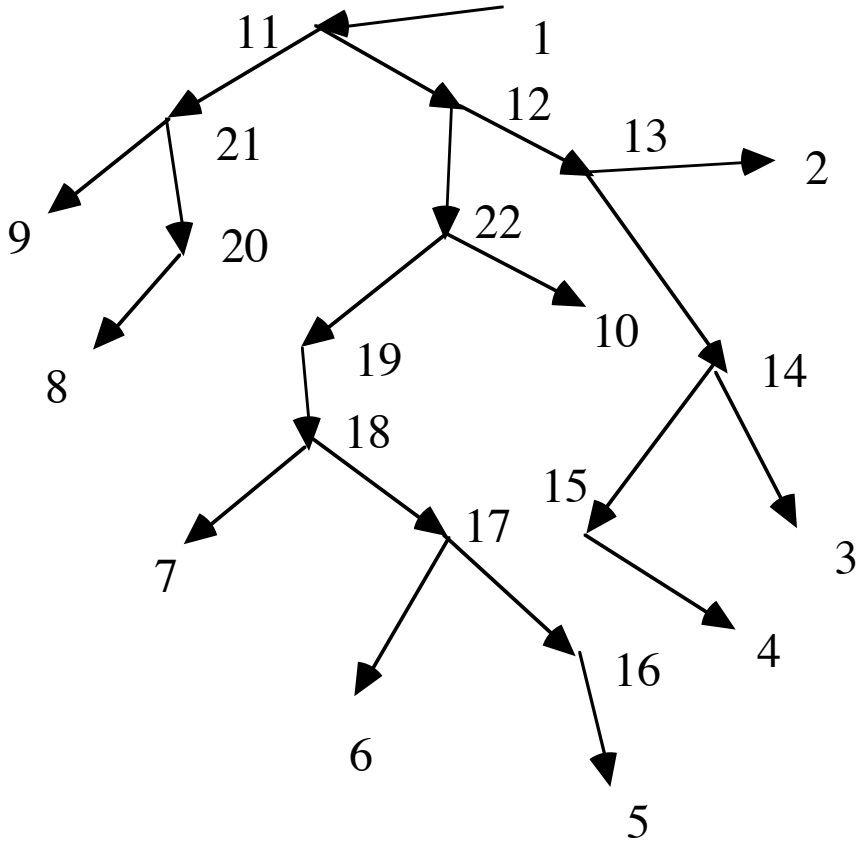

11
1
12
21
13
2
9
20
22
10
8
19
14
18
15
17
3
7
4
16
6
5

A **parent map** p is a map p: V-{r} → V such that
for each v in V-{r}, p(v) is a parent of v.
The set of all parent maps p is P(N) or P.



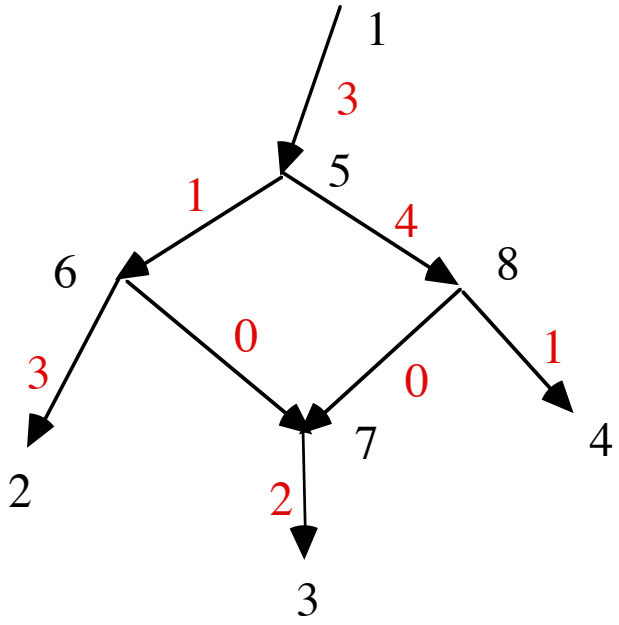
Each parent map p yields a displayed tree Np.

Example. If p(19)= 22 and p(16) = 17 then Np is



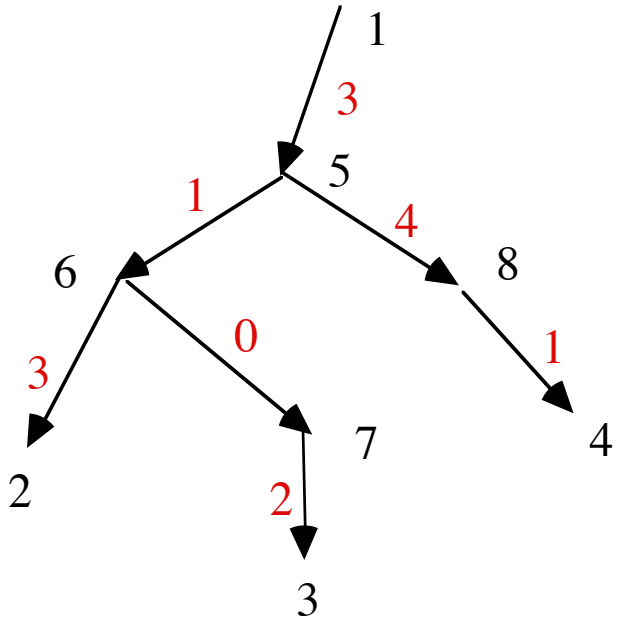
Main Definition. Suppose there are |P| parent maps p.
The **tree-average distance** for X on N is

$$d(x,y; N) = (1/|P|) \sum [d(x,y; Np): p \in P].$$



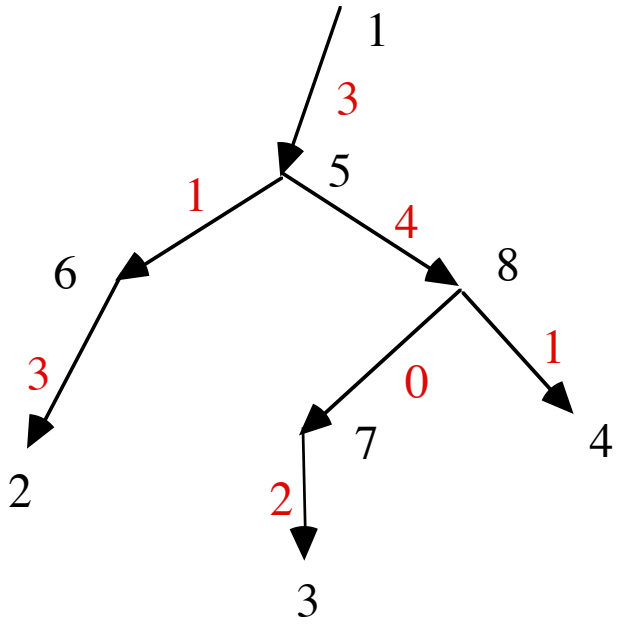
Example: Find d(1,2; N) and d(1,3; N).

Tree $T_1 = Np$ with $p(7) = 6$.



$d(1,2; T_1) = 7$
$d(1,3; T_1) = 6$

Tree T_2 = Np' with p'(7) = 8.



d(1,2; T_2) = 7
d(1,3; T_2) = 9

d(1,2; N) = (1/2) (7+7) = 7
d(1,3; N) = (1/2) (6+9) = 7.5

Theorem. Suppose that each tree-child arc has positive weight. Suppose each leaf is a tree-child. Then the tree-average distance d is a metric on X.

Major questions: **Suppose N is given and the tree-average distance is given on X.**

(1) When are the arc lengths uniquely determined?

(2) When is the topology of N uniquely determined?

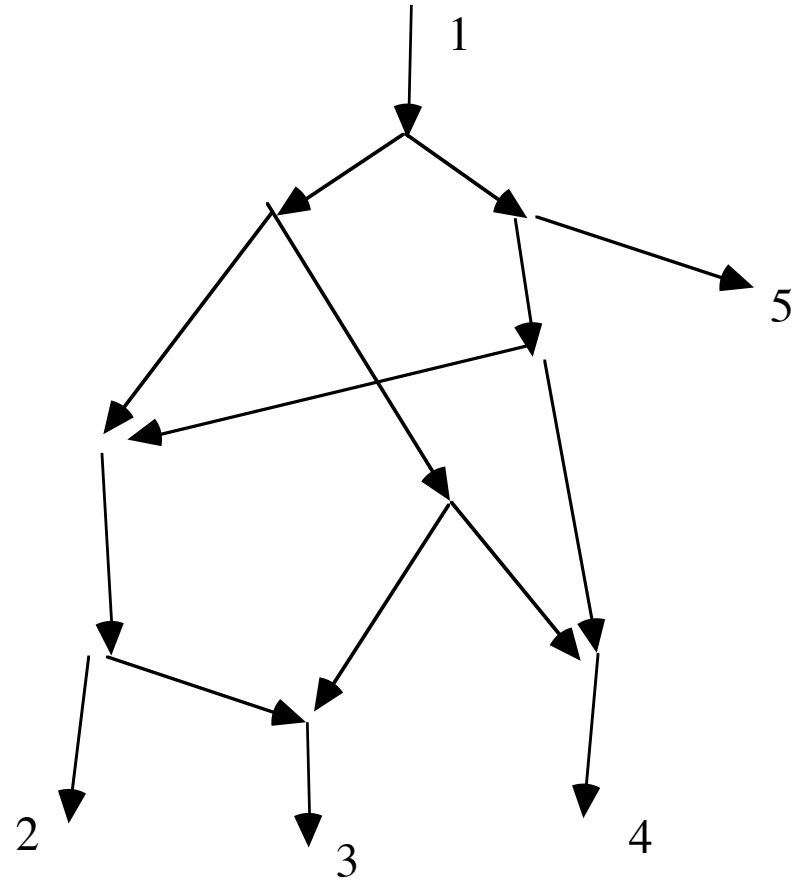
Major questions: **Suppose N is given and the tree-average distance is given on X.**

(1) When are the arc lengths uniquely determined?

For (1): If **d** is the vector of distances between members of X and **w** is the vector of arc-lengths of the normal arcs then one can find a matrix M so

$$\mathbf{d} = M\,\mathbf{w}.$$

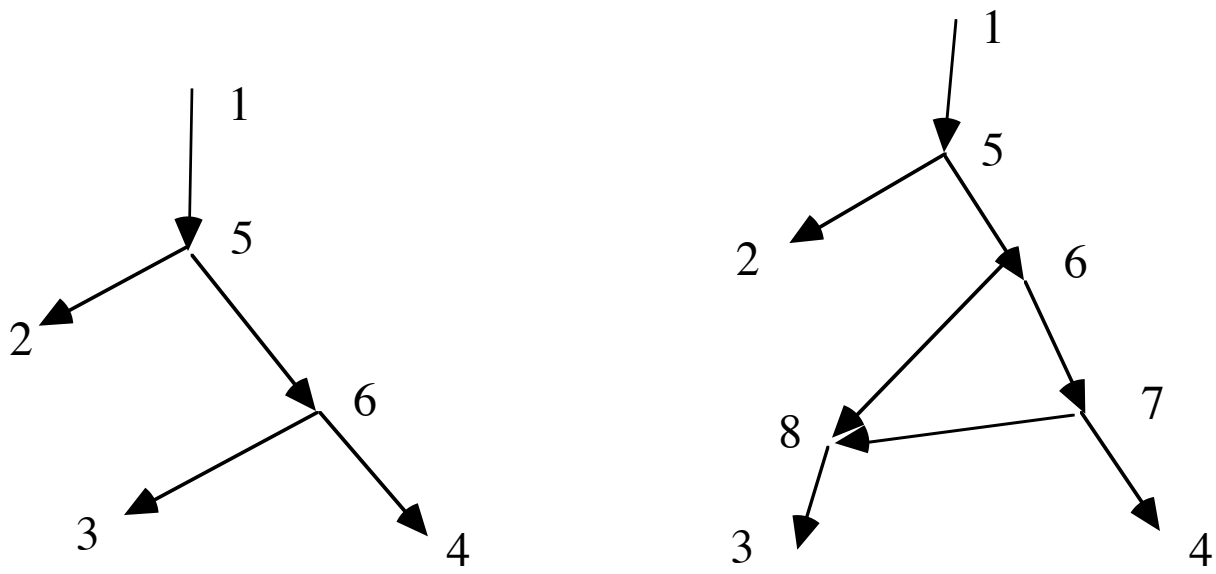
The arc lengths are uniquely determined provided that M has null-space 0.
For small examples, this can be worked, for example by using Mathematica.



There are 10 normal arcs and C(5,2) = 10 distances. But the arc lengths are NOT uniquely determined. (There is a relationship involving 7 arc lengths.)

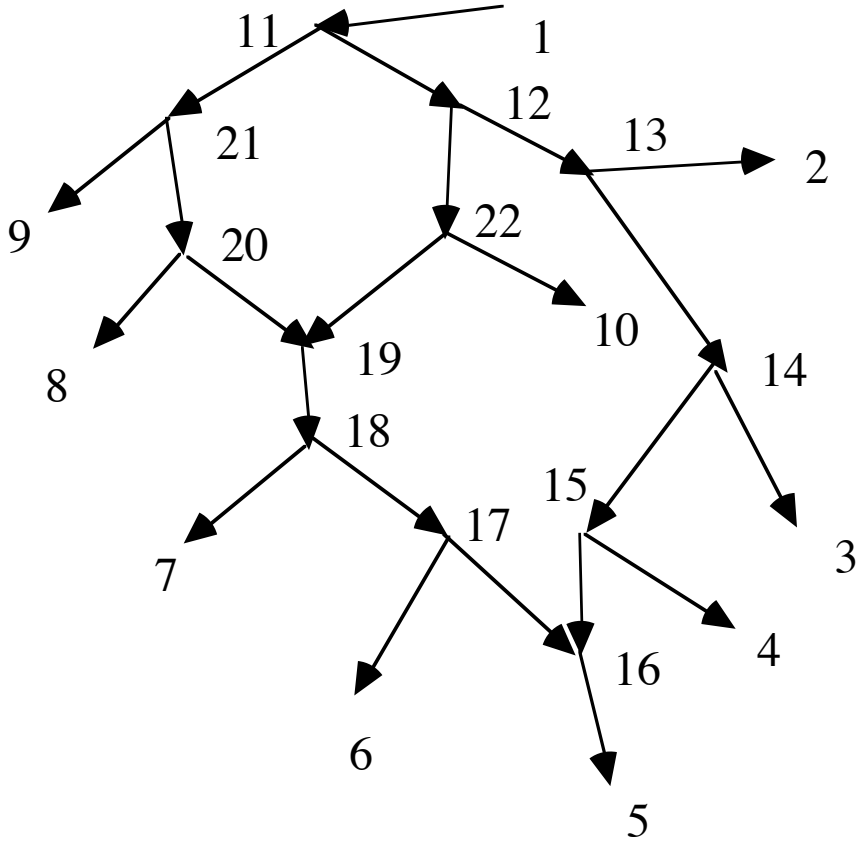
An arc (u,v) is **redundant** if there is a directed path from u to v different from the arc.
Redundant arcs can be problematic:

Start with a tree and add a redundant arc:

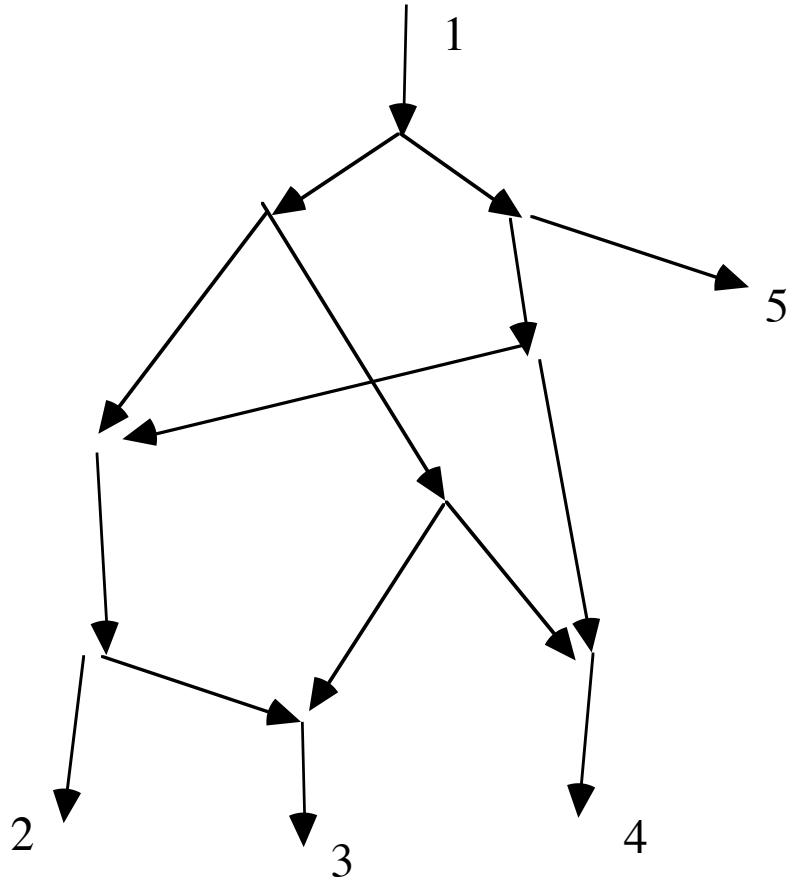


Now the arc lengths in N may not be uniquely determined (even though there are enough parameters-- 6 normal arcs and 6 distances).

A network is **normal** for X provided
(1) There are no redundant arcs.
(2) If v ∈ X, there exists a child c of v with indegree 1 (a **normal child**).



A normal network.



A network that is not normal.

Normal networks are fairly simple:

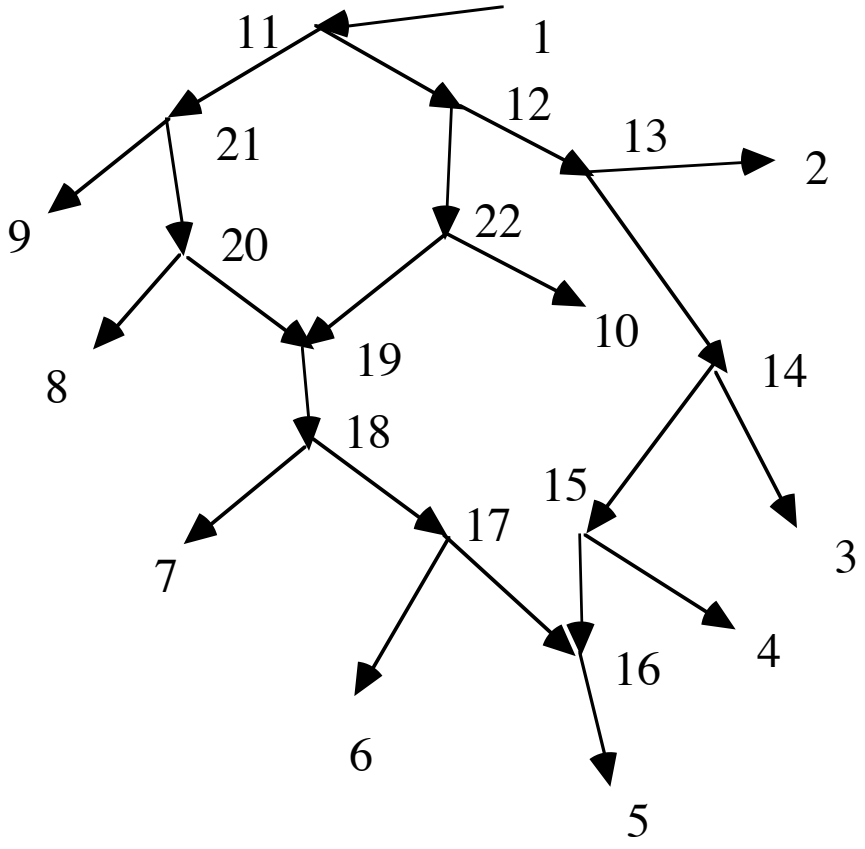
If N is normal and there are n tips, then
(1) There are at most n-2 hybrid vertices.
(2) There are at most $[n^2 + n - 2] / 2$ vertices.

Theorem. If N is normal then the trees Np are all distinct topologically.

Let Tr(N) be the set of trees with leaf set X displayed by N.

Corollary. $d(x,y; N) = (1/|Tr(N)|) \sum [d(x,y; T):$ N displays T$]$.

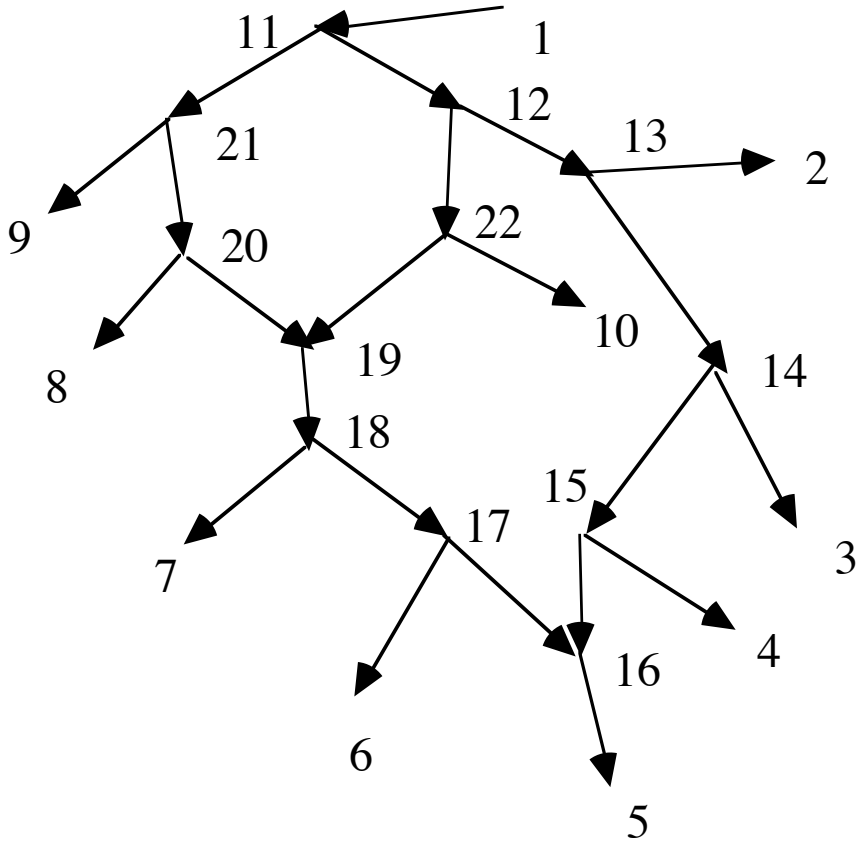
Main Theorem. Suppose N = (V,A,r,X) is normal for X. Assume
(1) All hybrids have indegree 2 and outdegree 1.
(2) Every weight of an arc to a hybrid vertex is 0.
Suppose that N is known and the tree-average distance d is known between members of X. Then all weights are uniquely determined.



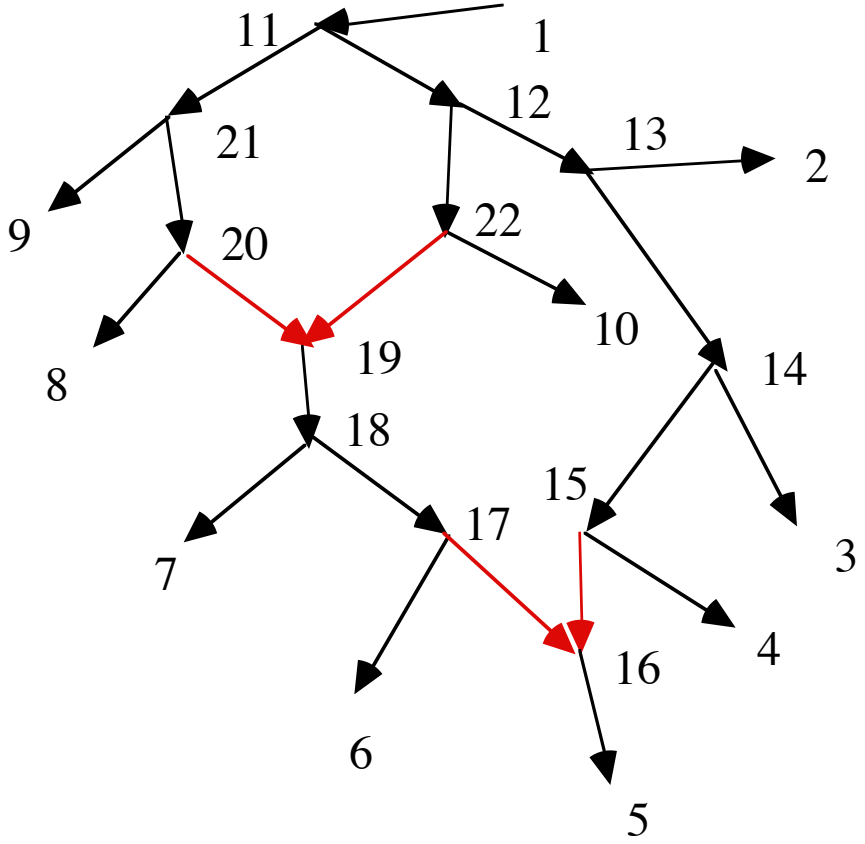
Corollary. Suppose |X| = n. Then the total number of arcs which lead into a normal vertex is at most C(n,2).

Moreover, the formulas are simple enough to give hope to unique reconstruction of the network as well.

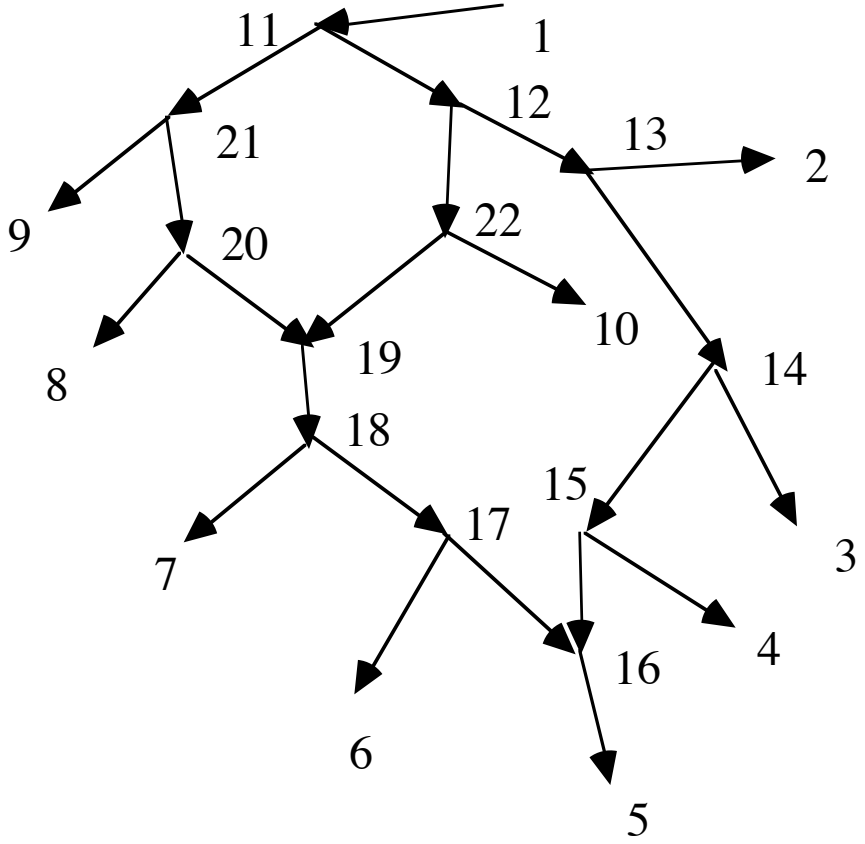
Conjecture: Suppose N = (V,A,r,X) is normal for X. Assume
(1) All hybrids have indegree 2 and outdegree 1.
(2) Every weight of an arc to a hybrid vertex is 0.
(3) The weight of every arc to a normal vertex is positive.
(4) All vertices have outdegree 0, 1, or 2.
(5) N has no pseudocycles.
Suppose that the tree-average distance d is known between all members of X. Then N is uniquely determined and can be constructed from d in polynomial time.


11
1
12
21
13
2
9
20
22
10
8
19
14
18
15
17
3
7
4
16
6
5

Some highlights of the proof of Theorem 1:



A n**ormal path from v to x** is a directed path starting at v such that each vertex after v is normal.

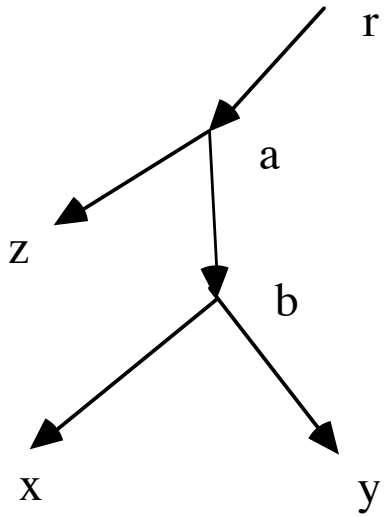


19,18,7 is normal
6 is normal
18,17,16,5 is not normal

Every arc on a normal path lies in Np for every parent map p.

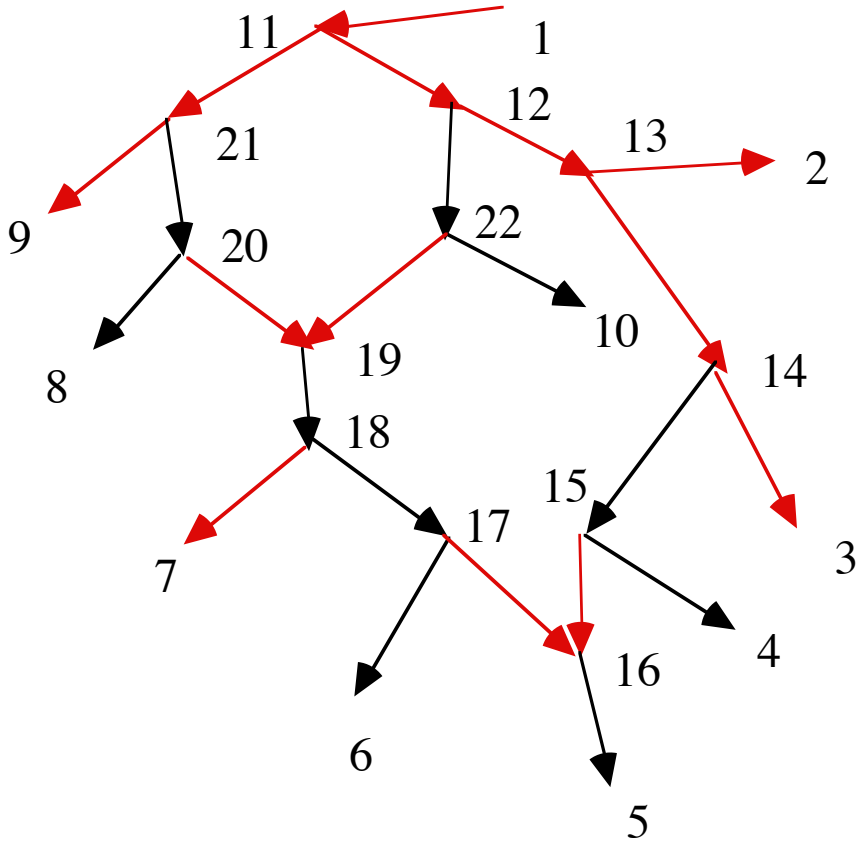
Rule 1. Consider the arc (a,b). Assume there is a normal path P from b to x in X, a normal path from b to y in X that is disjoint from P except at b, and a normal path from a to z in X that does not include b. Then

$$w(a,b) = [d(r,x)+d(z,y)-d(r,z)-d(x,y)]/2.$$



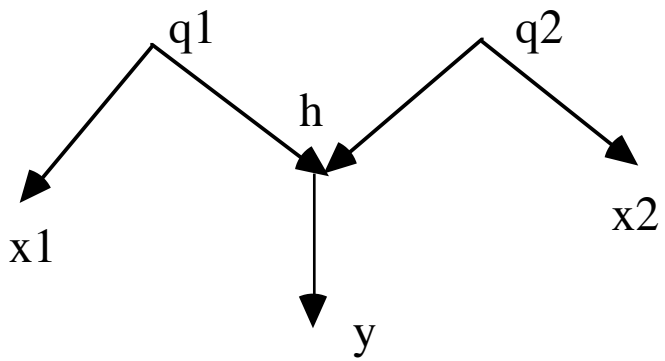
Proof. This is true in each tree Np.

There is a simpler formula if b is a normal leaf or a = r.



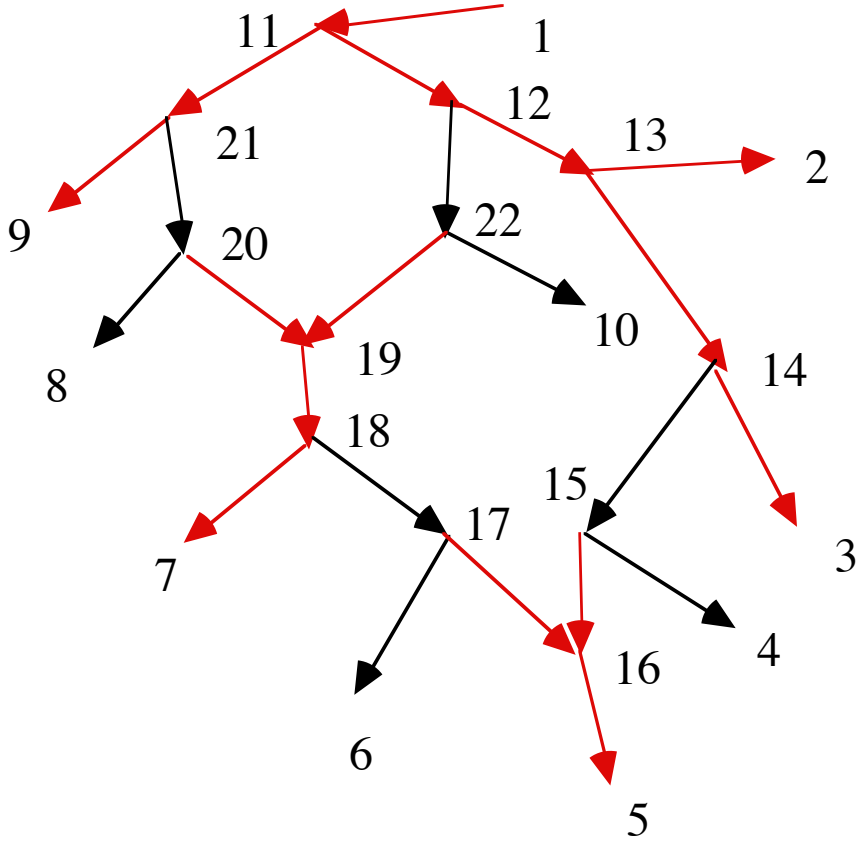
$$w(13,14) = [d(1,3) + d(2,4) - d(1,2) - d(3,4)] / 2$$

Rule 2. Suppose h is hybrid with parents q1 and q2 and child y. Choose a normal path from q1 to x1 in X, from q2 to x2 in X. Then
$w(h,y) = [d(y,x1)+d(y,x2)-d(x1,x2)\ ] / 2$



The proof uses that $w(q1,h) = w(q2,h) = 0$ and that the indegree of h is 2.

There is a related formula if instead there are two disjoint normal paths from the child of h.



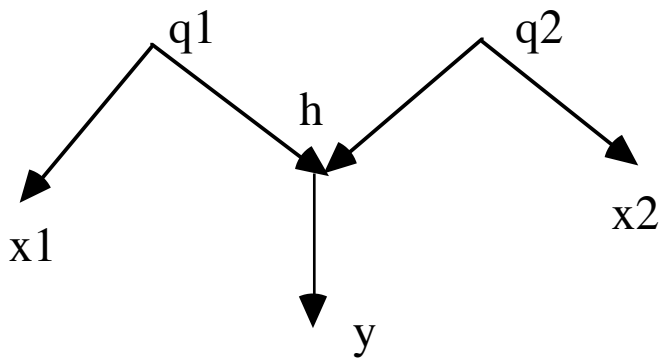
w(16,5) = [d(5,4)+d(5,6)-d(4,6)]/2

Rule 3. Suppose h is hybrid with parents q1 and q2. Suppose there are normal paths from q1 to x1 in X, from q2 to x2 in X, and from h to y in X. Then the length of the path from q1 to x1 is

$$d(x1,y) - d(r,y) + [d(r,x1)+d(r,x2)-d(x1,x2)] / 2.$$

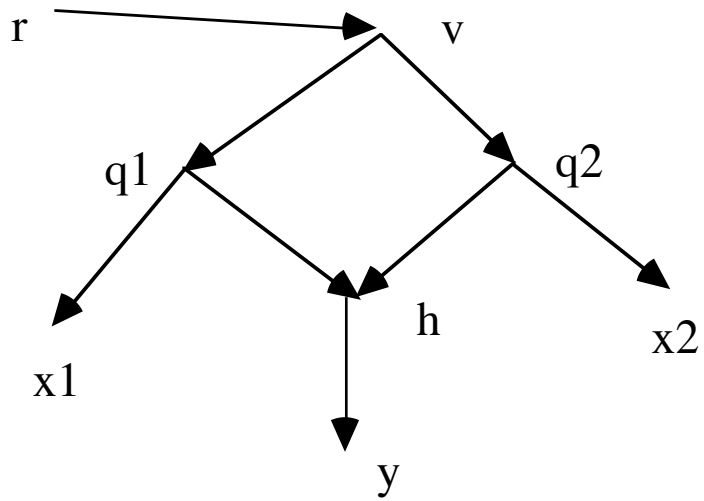
In particular, if x1 is a child of q1 then

$$w(q1,x1) = d(x1,y) - d(r,y) + [d(r,x1)+d(r,x2)-d(x1,x2)] / 2.$$



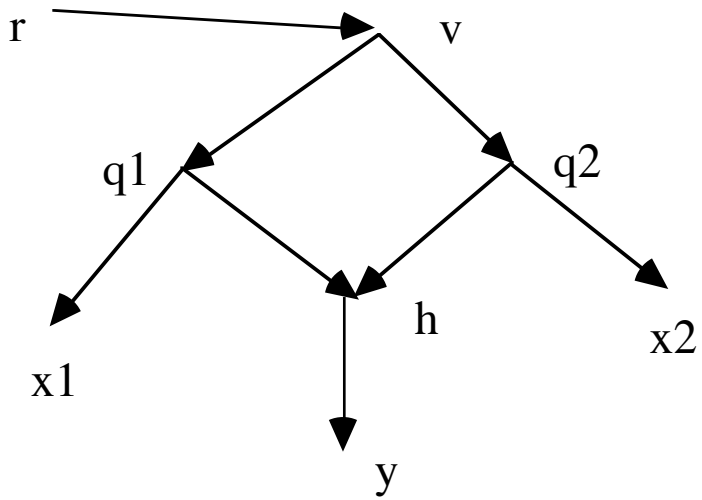
Idea of proof:
For any parent map p such that p(h) = q1 the **complementary parent map** p' agrees with p except that p(h) = q2.

Then Np ∪ Np' consists of Np together with (q2,h).



Part of Np ∪ Np'

Let C(N)
= d(x1,y;N) - d(r,y; N) + [d(r,x1; N)+d(r,x2; N)-d(x1,x2; N)] / 2

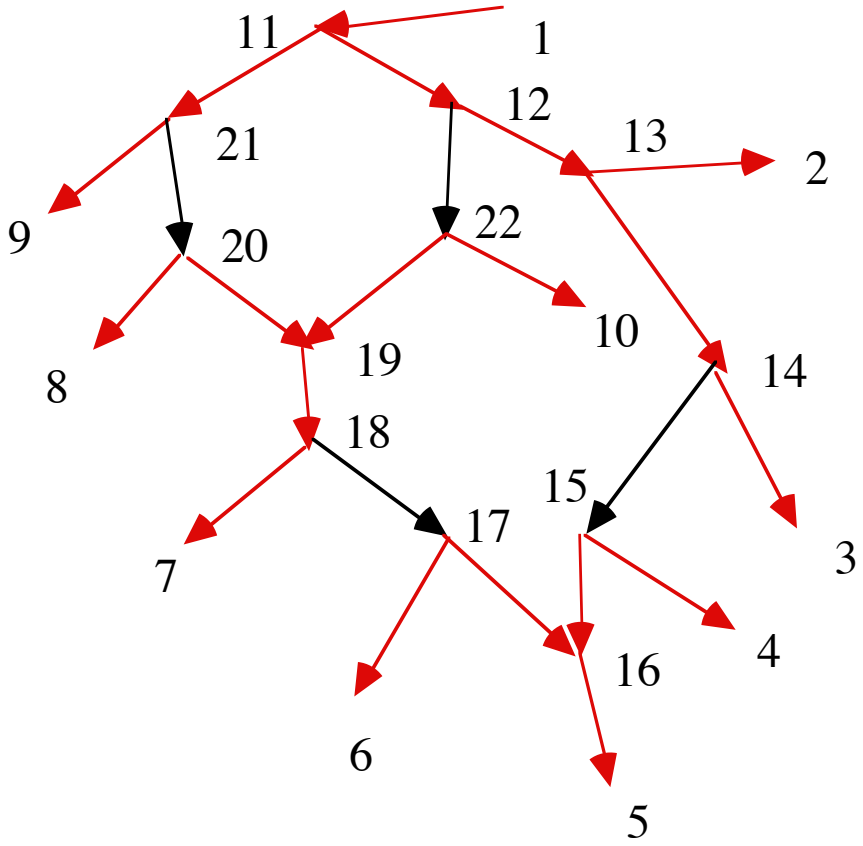


C(Np) = w(q1,x1) - d(v,q1; Np)

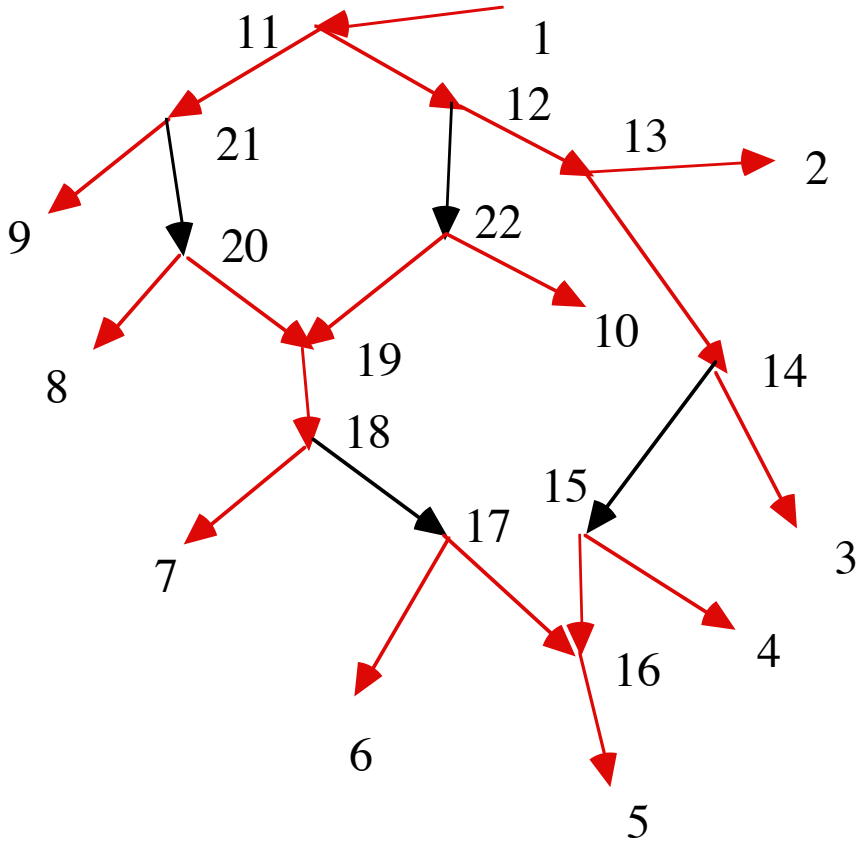
C(Np') = w(q1,x1) + d(v,q1; Np')

Hence C(Np) + C(Np') = 2 w(q1,x1)

C(N) = w(q1,x1)

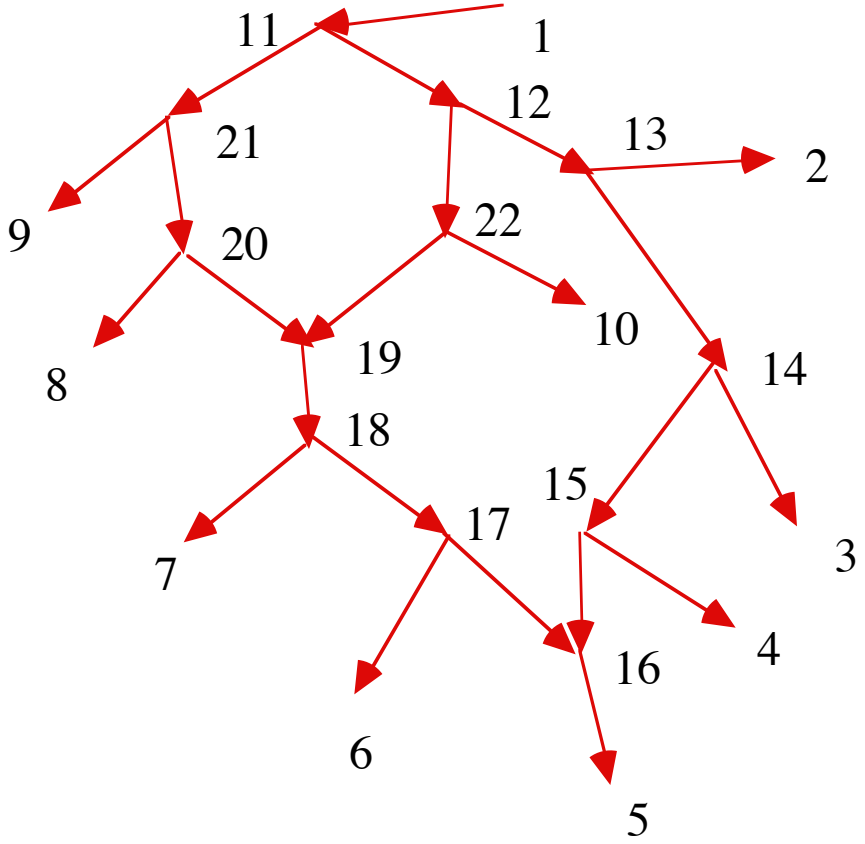


$$w(17,6) = d(6,5) - d(1,5) + [d(1,6)+d(1,4)-d(4,6)] / 2$$

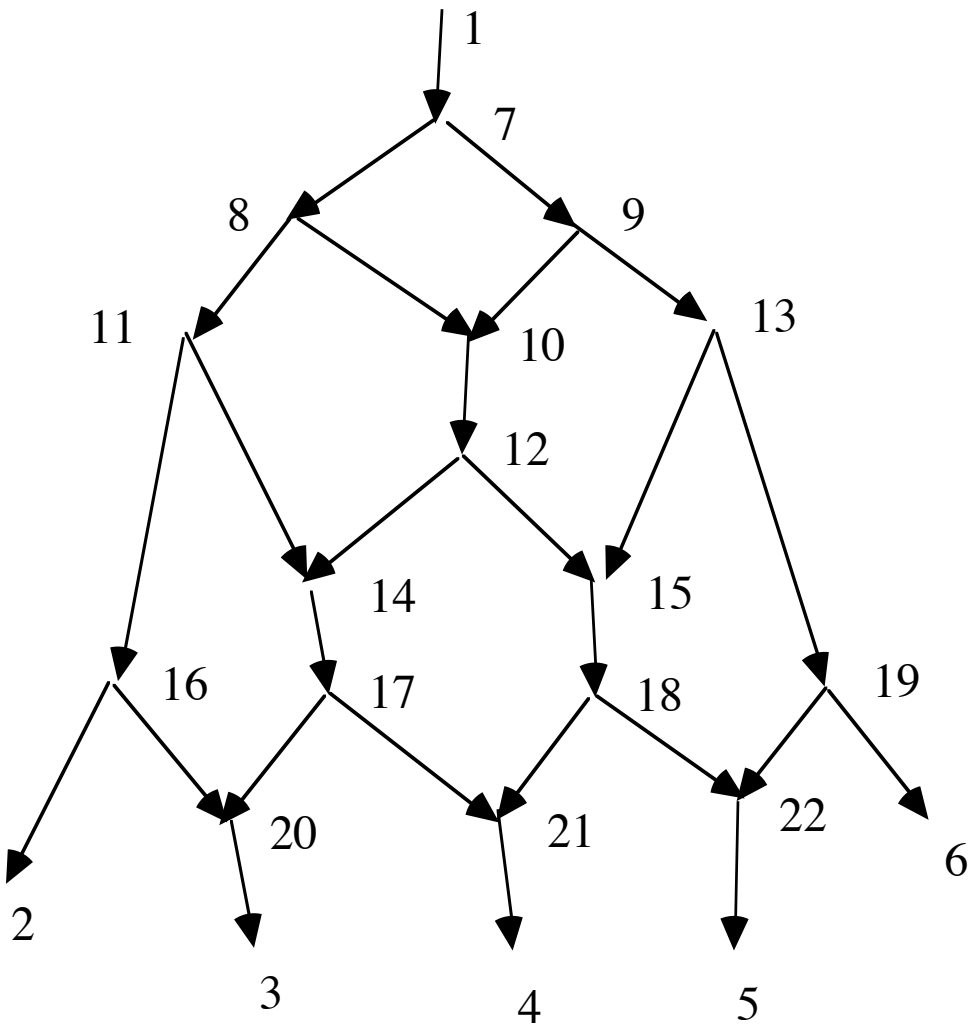


Since the path 21, 20, 8 is normal
d(21,8) = w(21,20)+w(20,8)
so w(21,20) = d(21,8) - w(20,8)
But d(21,8) = [d(8,1)+d(8,9)-d(1,9)]/2
Hence w(21,20) = [d(8,1)+d(8,9)-d(1,9)]/2 - w(20,8).

All the weights have been computed in this example.



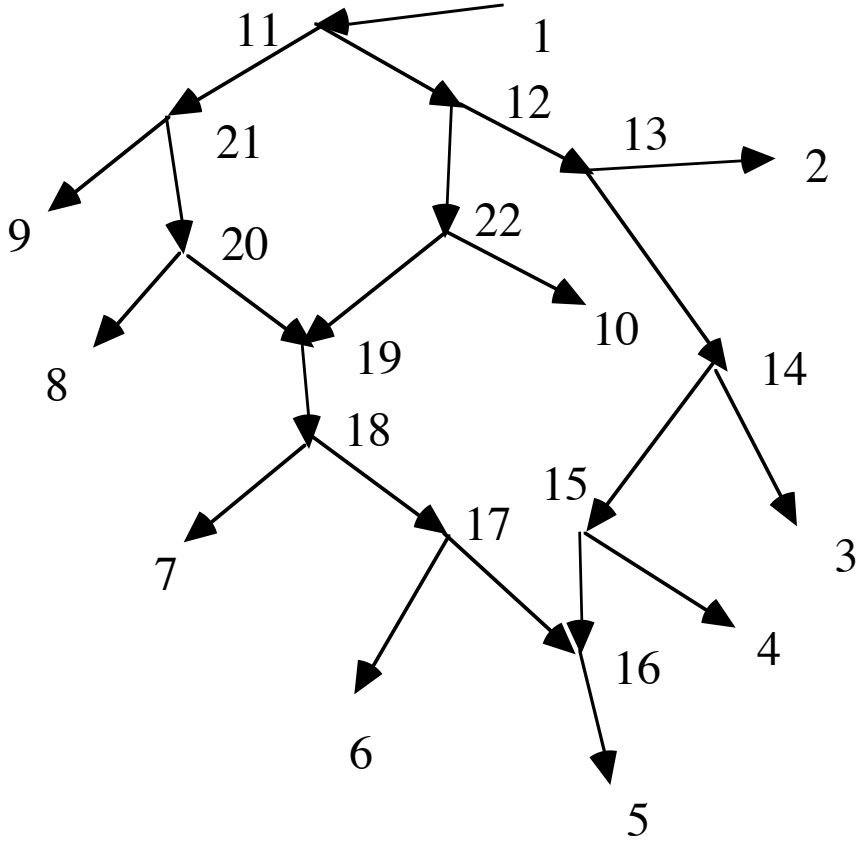
If the network is not normal, then often the arc lengths can be uniquely determined, yet the formulas are complicated and not "local."



w(20,3) =
(1/128) d(2,3) - (1/64) d(2,4)+ (1/64) d(2,5) - (1/128) d(2,6)
+ (1/64) d(3,4) - (1/64) d(3,5) + (1/128) d(3,6)

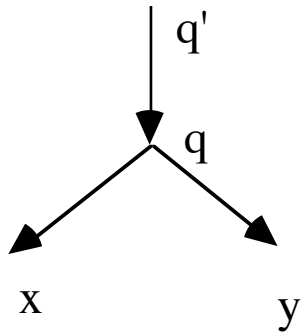
Conjecture: Suppose N = (V,A,r,X) is normal for X. Assume
(1) All hybrids have indegree 2 and outdegree 1.
(2) Every weight of an arc to a hybrid vertex is 0.
(3) The weight of every arc to a normal vertex is positive.
(4) All vertices have outdegree 0, 1, or 2.
(5) N has no pseudocycles.
Suppose that the tree-average distance d is known between all members of X. Then N is uniquely determined and can be constructed from d in polynomial time.

Status: There are lots of partial results and a working computer program.



Idea: A recursive construction can be performed with two basic steps:

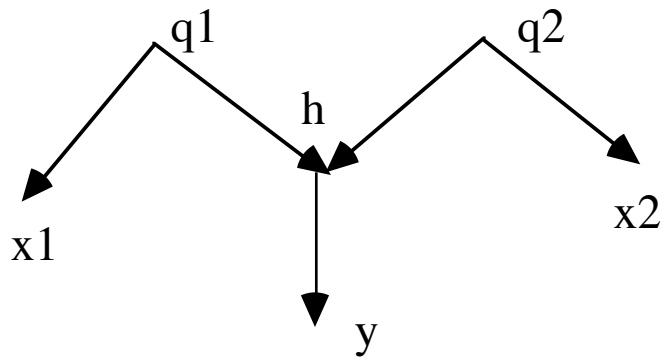
Step 1. Suppose {x,y} form a cherry with parent q.



By Rule 1 we know w(q,x) and w(q,y). Moreover, for every leaf z other than x or y, d(z,q) = d(z,x) - w(q,x) = d(z,y)- w(q,y). Hence we can remove x and y and insert q.

Moreover, there is a necessary and sufficient condition in terms of the tree-average distance to recognize a cherry {x,y}.

Step 2. Suppose y is a tree-child of a hybrid and we know tree-children of the parents. We know y, x1, x2.



By Rule 2 we compute w(h,y). By Rule 3 we compute w(q1,x1) and w(q2,x2).
For all leaves z other than y, x1, x2
d(z,q1) = d(z,x1) - w(q1,x1)
d(z,q2) = d(z,x2) - w(q2,x2)
d(q1,q2) = d(x1,x2) - w(q1,x1) - w(q2,x2).
We can remove x1,y, x2, h and insert q1 and q2.

Moreover, there are strong necessary conditions in terms of the tree-average distance to recognize this situation.

Future problems:

(1) Which practical distances behave sufficiently like the tree-average distance?

(a) Take an average of the distances on gene trees for various genes using some standard distances such as Jukes-Cantor or the log-determinant distance.

(b) Modify some standard distances such as Jukes-Cantor or the log-determinant distance.

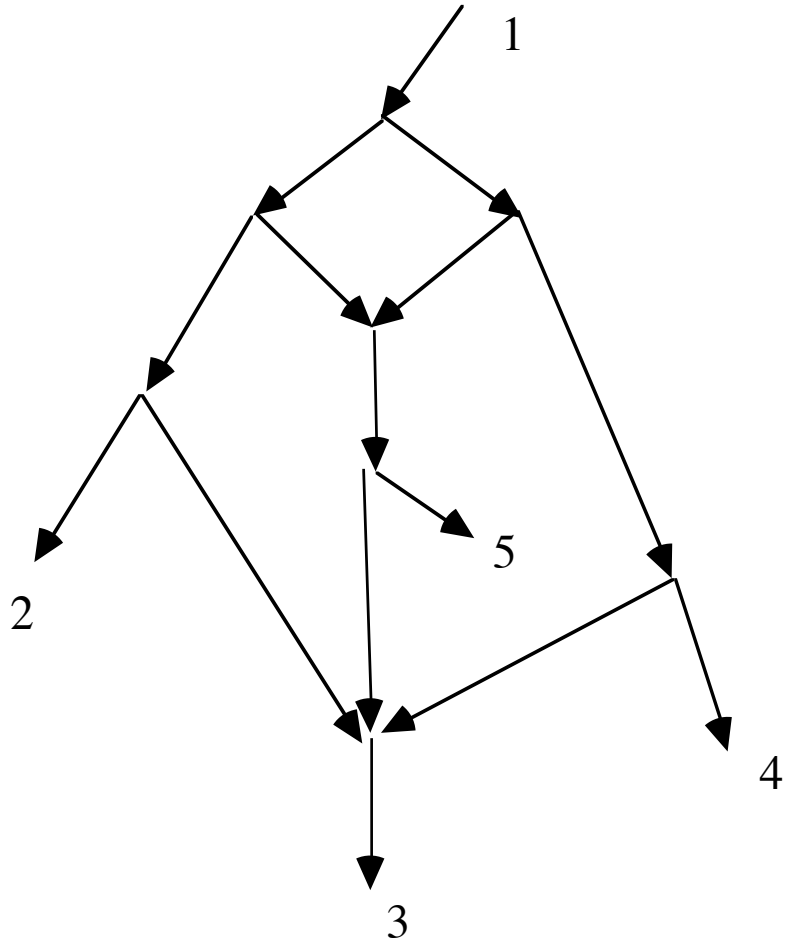
(2) Find more robust, practical formulas for the weights of arcs that are don't require exact values.

(a) Errors in formulas like
$d(x1,y) - d(r,y) + [d(r,x1)+d(r,x2)-d(x1,x2)] /2$

propagate to create larger errors in later approximations during recursive reconstruction of the network.

(b) Accuracy in the distances from the root r are especially crucial.

(3) Deal with hybrids of indegree 3 or higher.



Summary

1. To use distance methods on phylogenetic networks N, we need to restrict the kinds of networks allowed.

2. The tree-average distance between x and y is the average of the distances between x and y in the various displayed trees Np.

3. Given a normal network N whose hybrids have indegree 2, and given the tree-average distance function on the leaves, we can find the weights of all arcs. Often we can find N itself from the tree-average distance function.

Thanks to the Isaac Newton Institute.
Thanks to the organizers.
Thank you for your attention.