Optimising phylogenetic diversity on phylogenetic networks

Charles Semple

School of Mathematics and Statistics
University of Canterbury, New Zealand

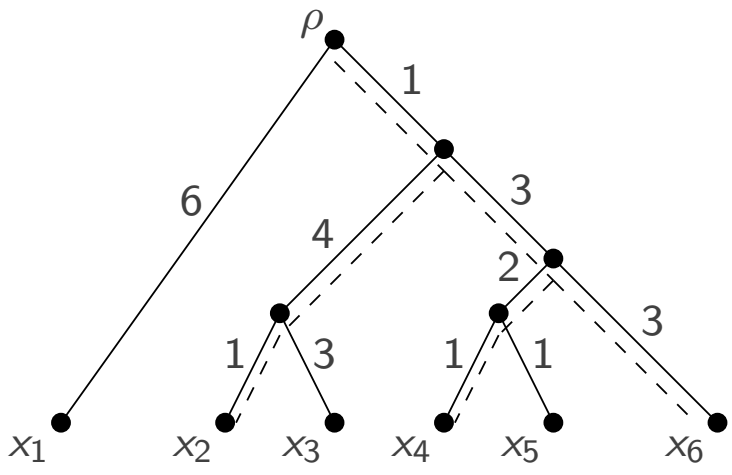
- ⋆ Phylogenetic diversity (PD) is a measure for quantifying the biodiversity of a set of present-day taxa (Faith, 1992).
- ⋆ PD quantifies the extent to which the taxa spans the 'Tree of Life'.
- ⋆ Features arise at a rate proportional to edge lengths, and no features are loss.

- ⋆ Phylogenetic diversity (PD) is a measure for quantifying the biodiversity of a set of present-day taxa (Faith, 1992).
- ⋆ PD quantifies the extent to which the taxa spans the 'Tree of Life'.
- ⋆ Features arise at a rate proportional to edge lengths, and no features are loss.

- ⋆ The underlying optimisation problem is to find, for a given set X of taxa and a positive integer k, a subset of X of size k that maximises PD.

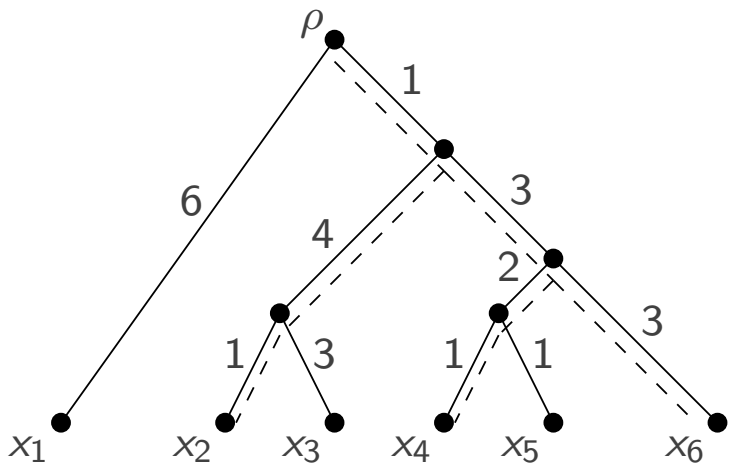
Phylogenetic diversity on trees

Let T be a (rooted) phylogenetic tree on X. The phylogenetic diversity of a subset S of X, denoted $PD(S)$, is the sum of the lengths of all edges on a path from the root of T to an element in S.



Phylogenetic diversity on trees

Let T be a (rooted) phylogenetic tree on X. The phylogenetic diversity of a subset S of X, denoted $PD(S)$, is the sum of the lengths of all edges on a path from the root of T to an element in S.



MAX-PD-TREE

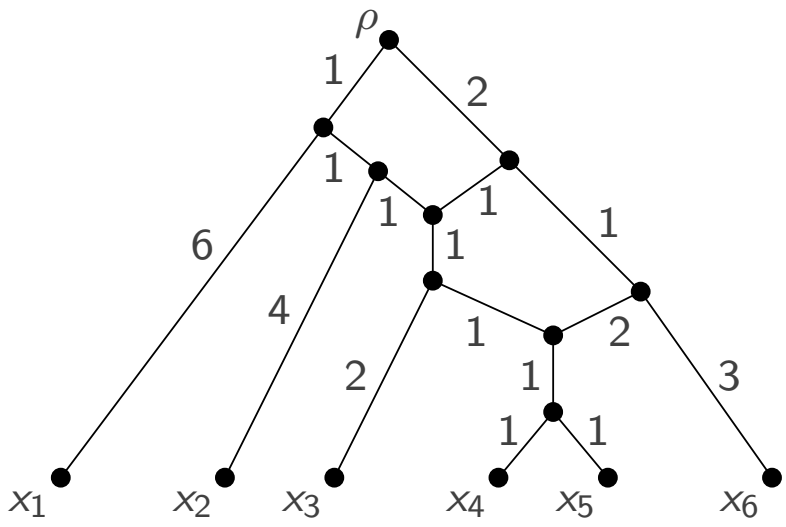
Input: A phylogenetic tree T on X and a positive integer k.

Objective: Find the maximum value of $PD(S)$ over all subsets S of X of size k.

- ⋆ Fast algorithms exist for solving MAX-PD-TREE (Pardi, Goldman, 2005; Steel, 2005).

Phylogenetic diversity on networks

Let N be a phylogenetic network on X. The phylogenetic diversity of a subset S of X is the sum of the lengths of all edges on a path from the root of N to an element in S.



MAX-PD-NETWORK

Input: A network N and a positive integer k.

Objective: Find the maximum value of $PD(S)$ over all subsets S of X of size k.

Complexity of Max-PD-Network

Theorem 1. (Bordewich, S, Wicke, 2022)

- ⋆ Max-PD-Network is NP-hard.
- ⋆ Unless $P = NP$, Max-PD-Network cannot be approximated in polynomial time with an approximation ratio better than $1 - \frac{1}{e}$. □

Complexity of Max-PD-Network

Theorem 1. (Bordewich, S, Wicke, 2022)

- ⋆ Max-PD-Network is NP-hard.
- ⋆ Unless $P = NP$, Max-PD-Network cannot be approximated in polynomial time with an approximation ratio better than $1 - \frac{1}{e}$. □

- ⋆ A network is normal if it has no sibling and no stack reticulations, and no shortcuts.
- ⋆ Max-PD-Network is NP-hard when restricted to the class of normal networks.

Complexity of Max-PD-Network

Theorem 1. (Bordewich, S, Wicke, 2022)

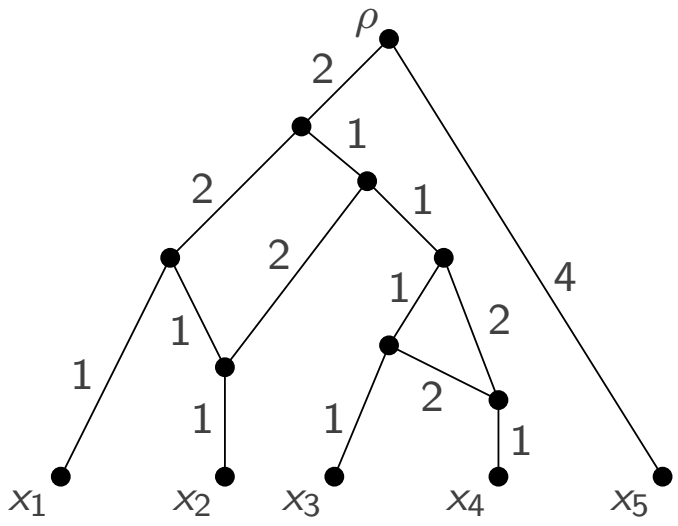
- ⋆ Max-PD-Network is NP-hard.
- ⋆ Unless $P = NP$, Max-PD-Network cannot be approximated in polynomial time with an approximation ratio better than $1 - \frac{1}{e}$. □

- ⋆ A network is normal if it has no sibling and no stack reticulations, and no shortcuts.
- ⋆ Max-PD-Network is NP-hard when restricted to the class of normal networks.

- ⋆ The function PD on 2^X, which assigns each subset S of X the value $PD(S)$, is a submodular function.
- ⋆ A fast algorithm exists that returns a $1 - \frac{1}{e}$ approximation for Max-PD-Network.

Level-one networks

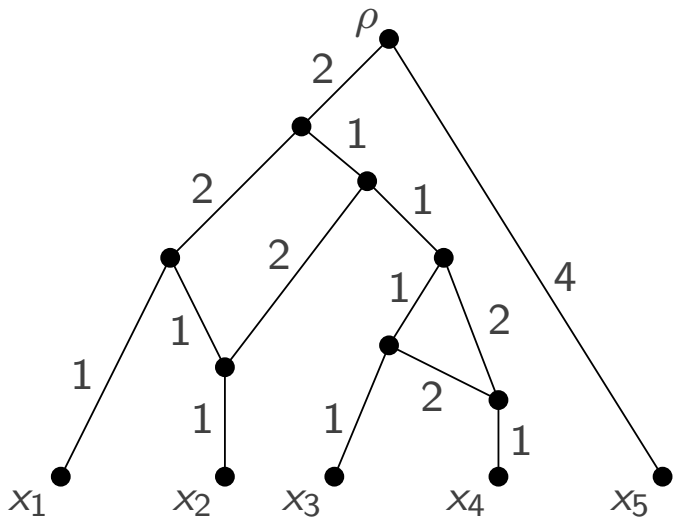
A network is level-1 if the underlying cycles are vertex disjoint.



Theorem 2. (Bordewich, S, Wicke, 2022) If N is a level-1 network, then Max-PD-Network is solvable in polynomial time in the size of X. □

Level-one networks

A network is level-1 if the underlying cycles are vertex disjoint.



Theorem 2. (Bordewich, S, Wicke, 2022) If N is a level-1 network, then Max-PD-Network is solvable in polynomial time in the size of X. □

- ⋆ Optimising $PD_{T_1}(S) + PD_{T_2}(S)$ over all subsets S of size k is the polynomial-time problem Weighted-Average-PD-on-2-Trees (Bordewich, S, Spillner, 2009).