Computing phylogenetic diversity for split systems

EMBO Workshop on current challenges and problems in phylogenetics

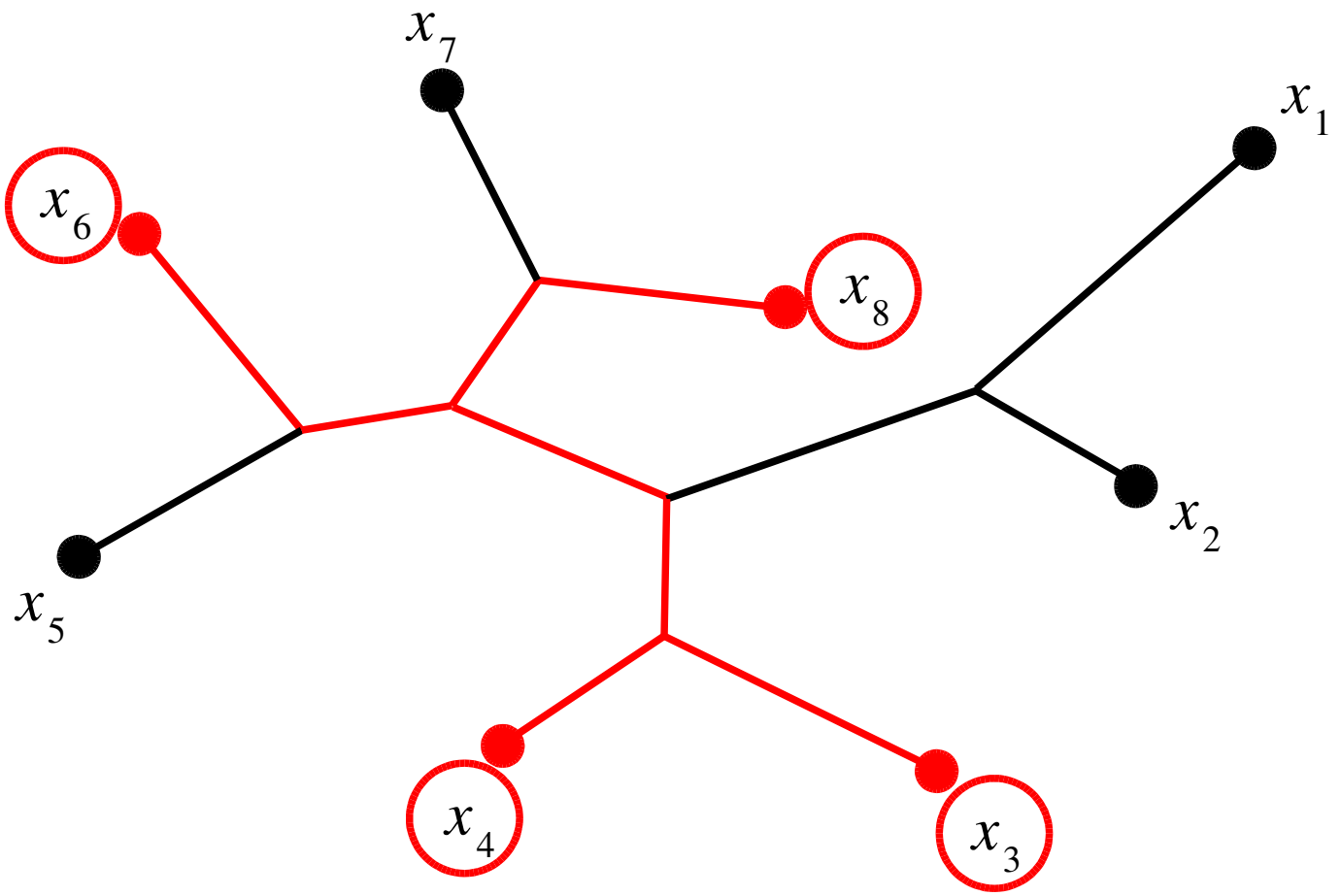
6 September 2007

Andreas Spillner Binh T. Nguyen Vincent Moulton

Outline

- **Preliminaries**
  - Definitions
  - Motivation
- **Computational complexity of the problem**
  - General split systems
  - Compatible split systems
  - Circular split systems
- **Summary**

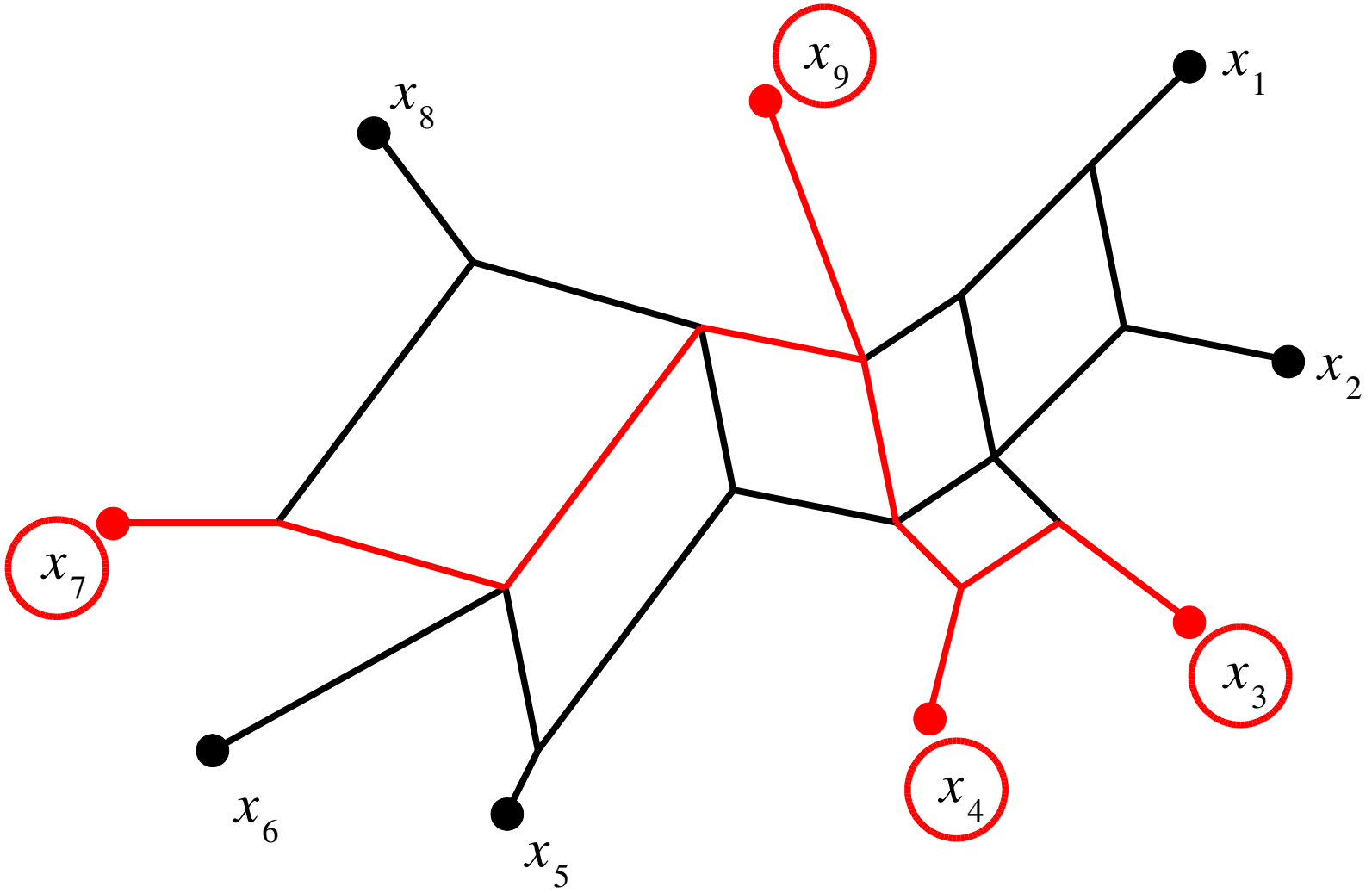
Phylogenetic diversity for trees



$PD(Y) =$ length of the shortest tree connecting Y

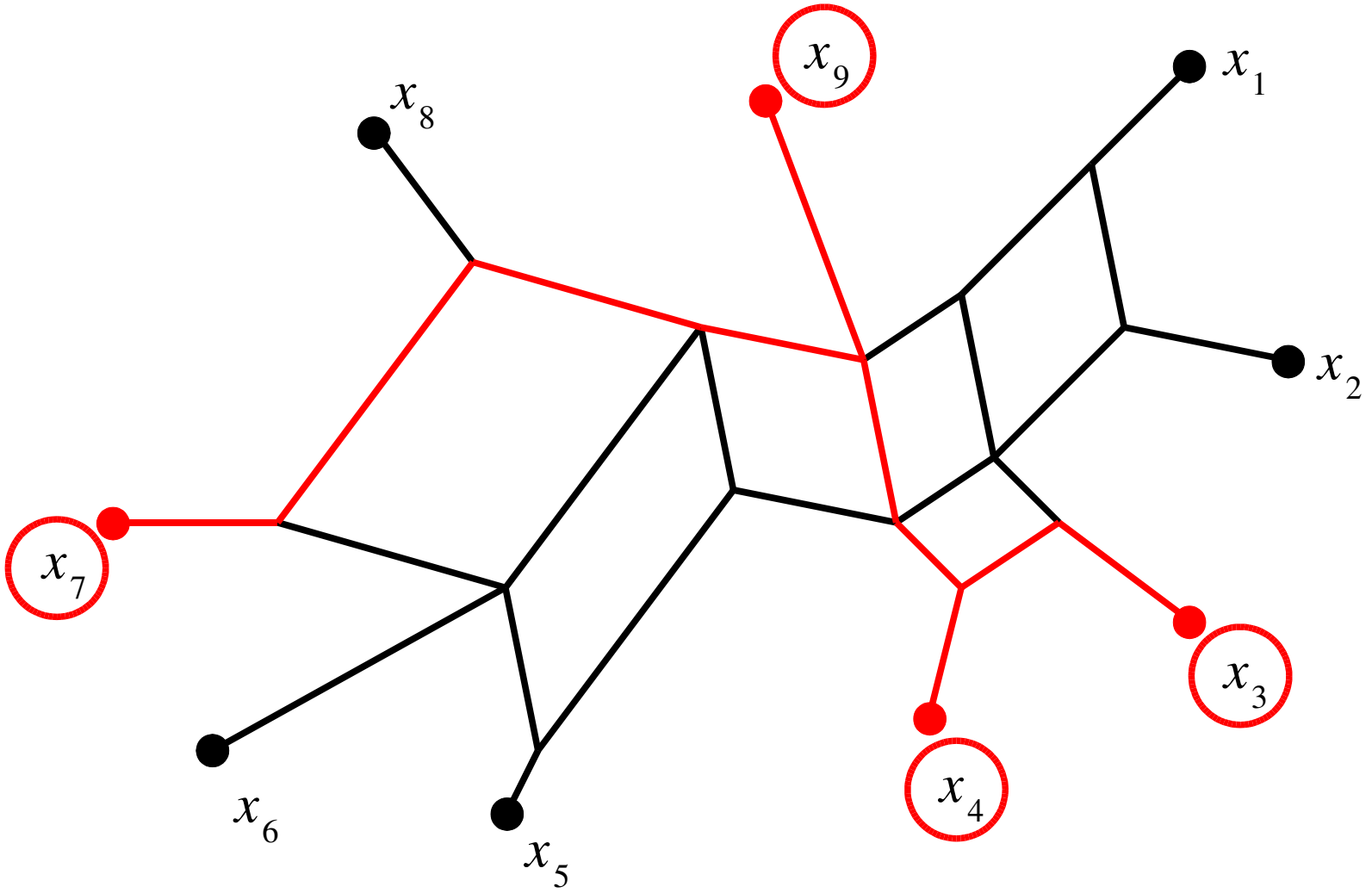
Faith, 1992

Phylogenetic diversity for networks



$PD(Y)$ = length of a shortest tree connecting Y

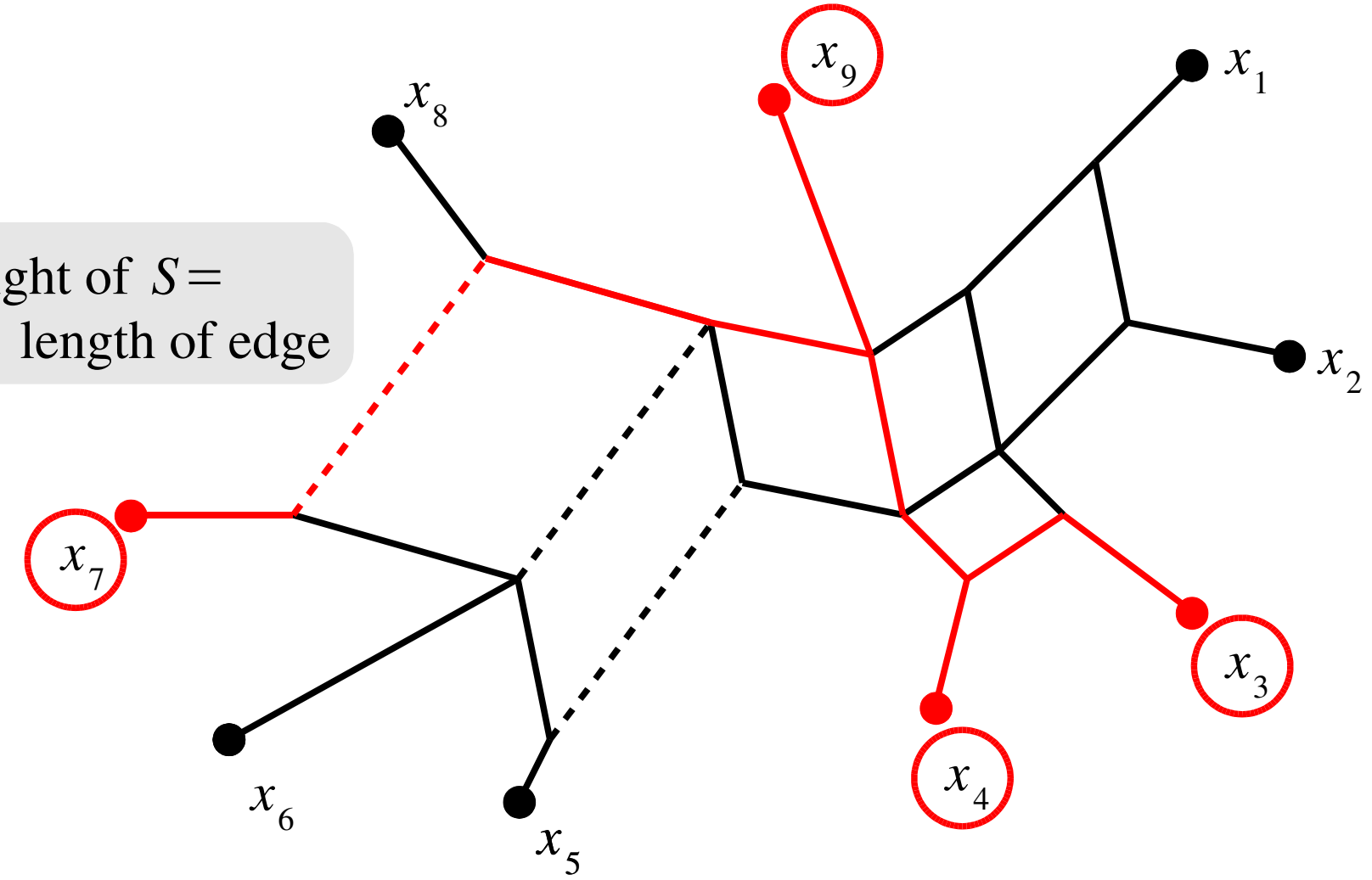
Phylogenetic diversity for networks



An alternative shortest tree connecting Y.

Phylogenetic diversity for split systems

weight of S = length of edge



split $S = x_5 x_6 x_7 | x_1 x_2 x_3 x_4 x_8 x_9$

Phylogenetic diversity for split systems

$\mathbf{S}$ ... a set of splits of X (split system)

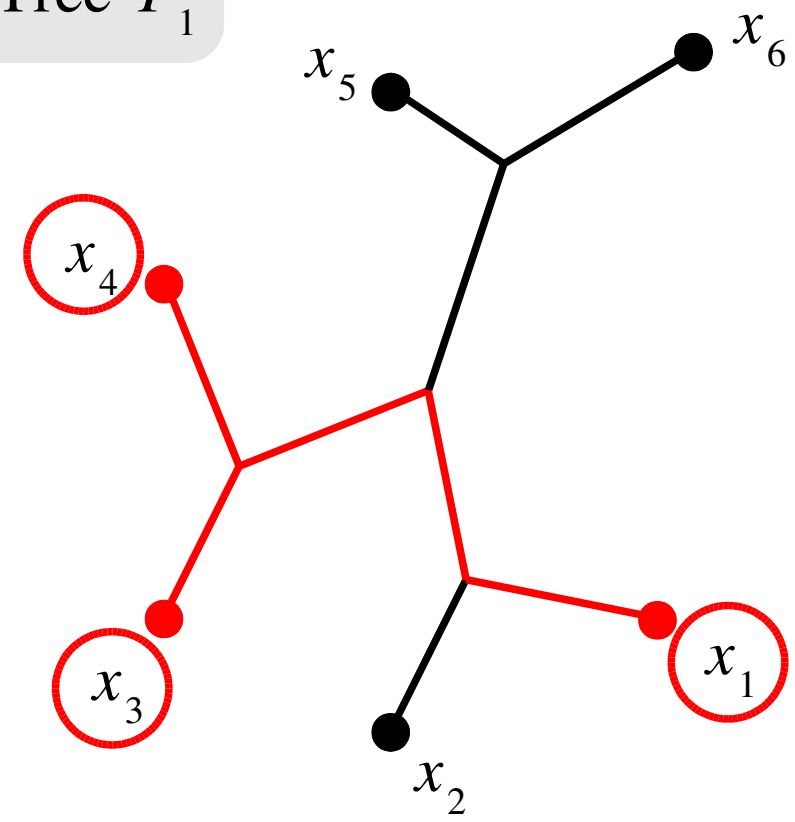
$\omega : \mathbf{S} \to \mathbb{R}_{\geqslant 0}$ ... split weigt function

$Y \subseteq X$ ... subset of X

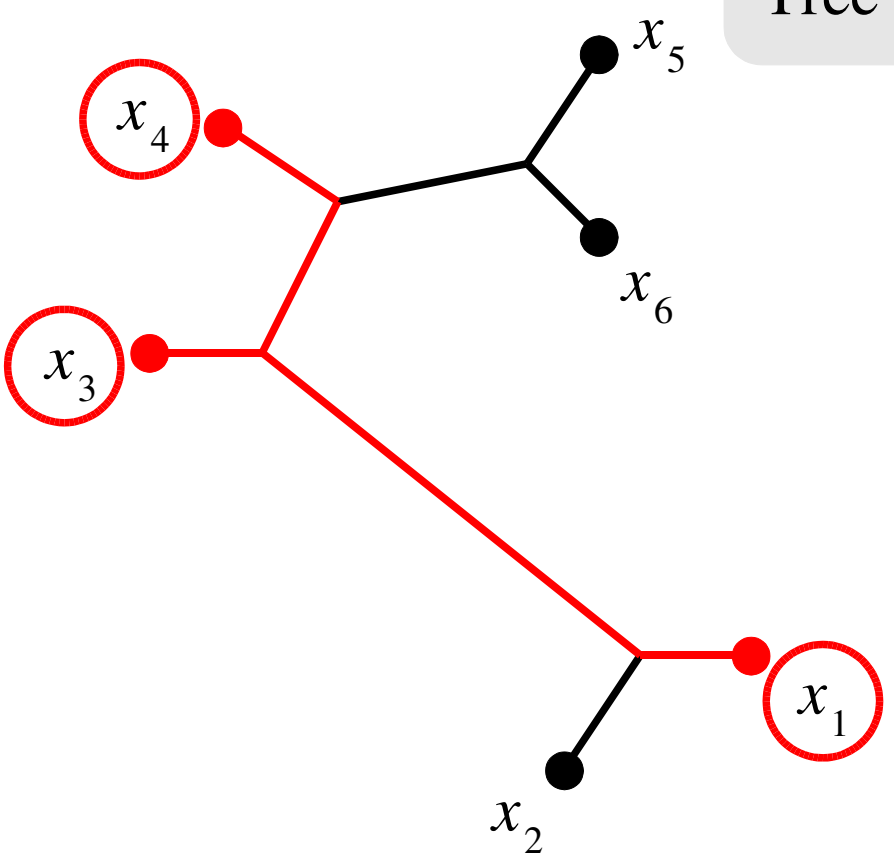
$$PD_{\mathbf{S}}(Y) = \sum_{\substack{A|B=S \in \mathbf{S} \\ A \cap Y \neq \emptyset \\ B \cap Y \neq \emptyset}} \omega(S)$$ ... phylogenetic diversity of Y

Motivation: Several trees on the same set of taxa

Tree T_1



Tree T_2

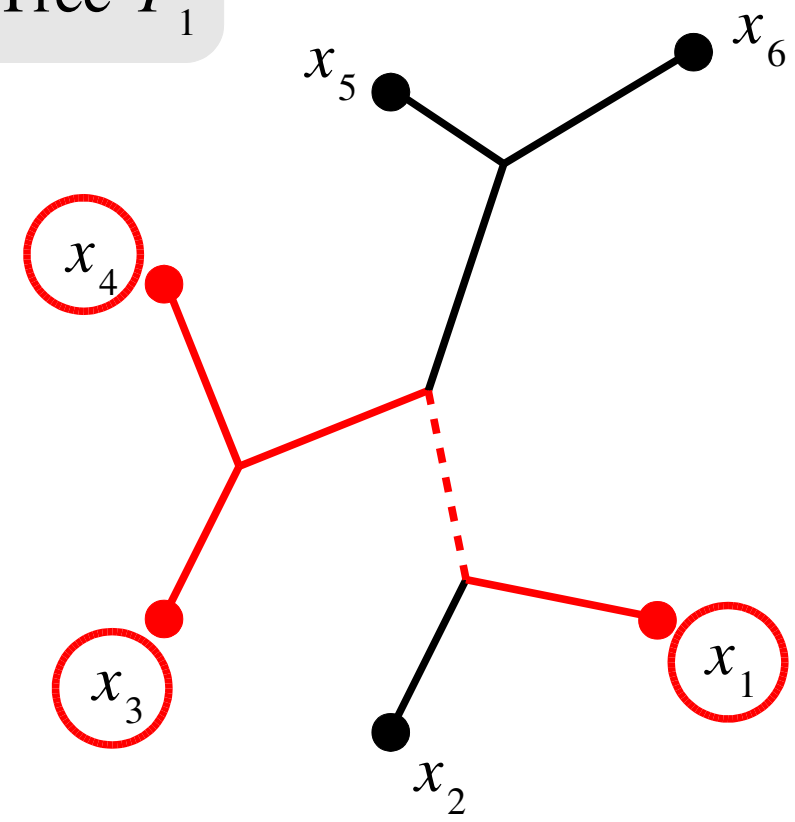


Goal: Maximize the sum of the lengths of red subtrees.

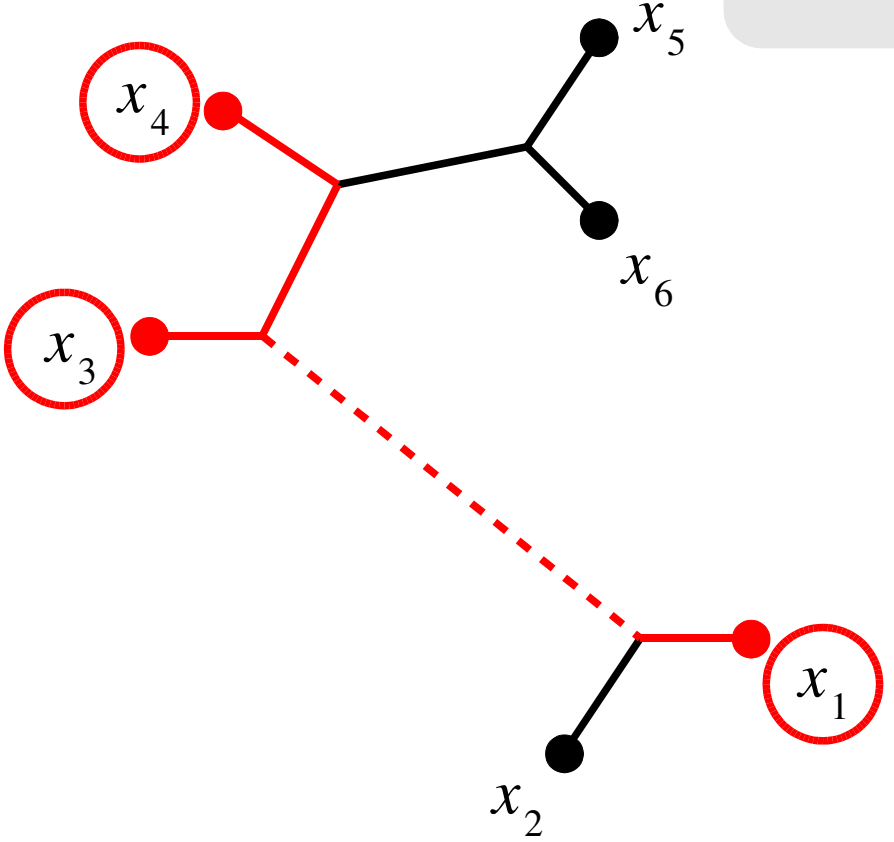
Minh et al., 2006

Motivation: Several trees on the same set of taxa

Tree T_1



Tree T_2



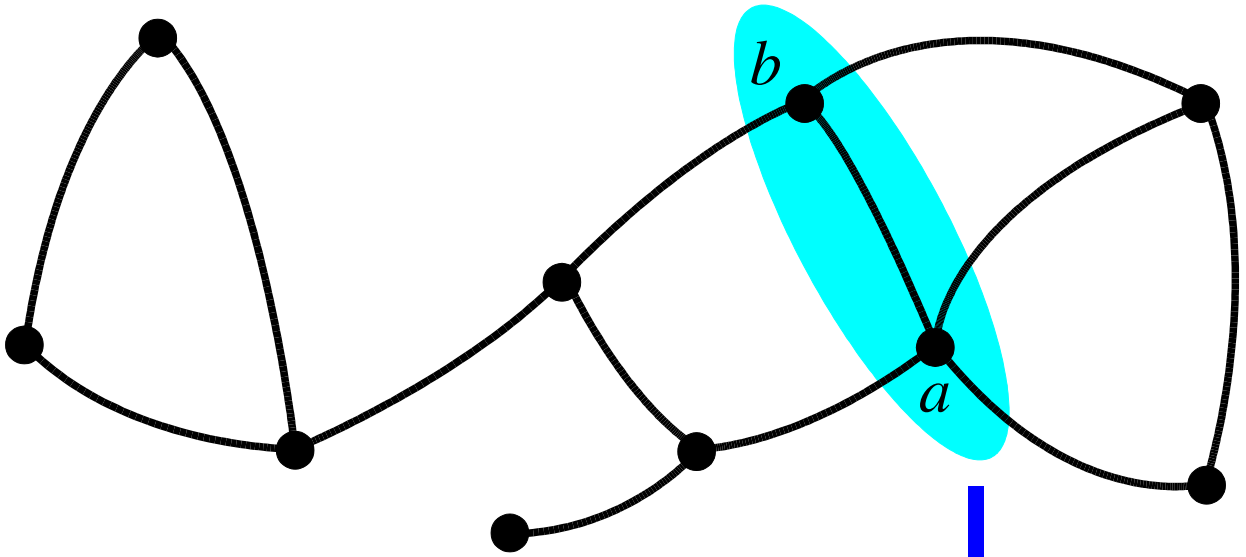
Represent the splits from both trees as a single split system.

Outline

- Preliminaries
  - Definitions
  - Motivation
- **Computational complexity of the problem**
  - General split systems
  - Compatible split systems
  - Circular split systems
- Summary

The general problem is NP-hard

Graph $G=(V,E)$

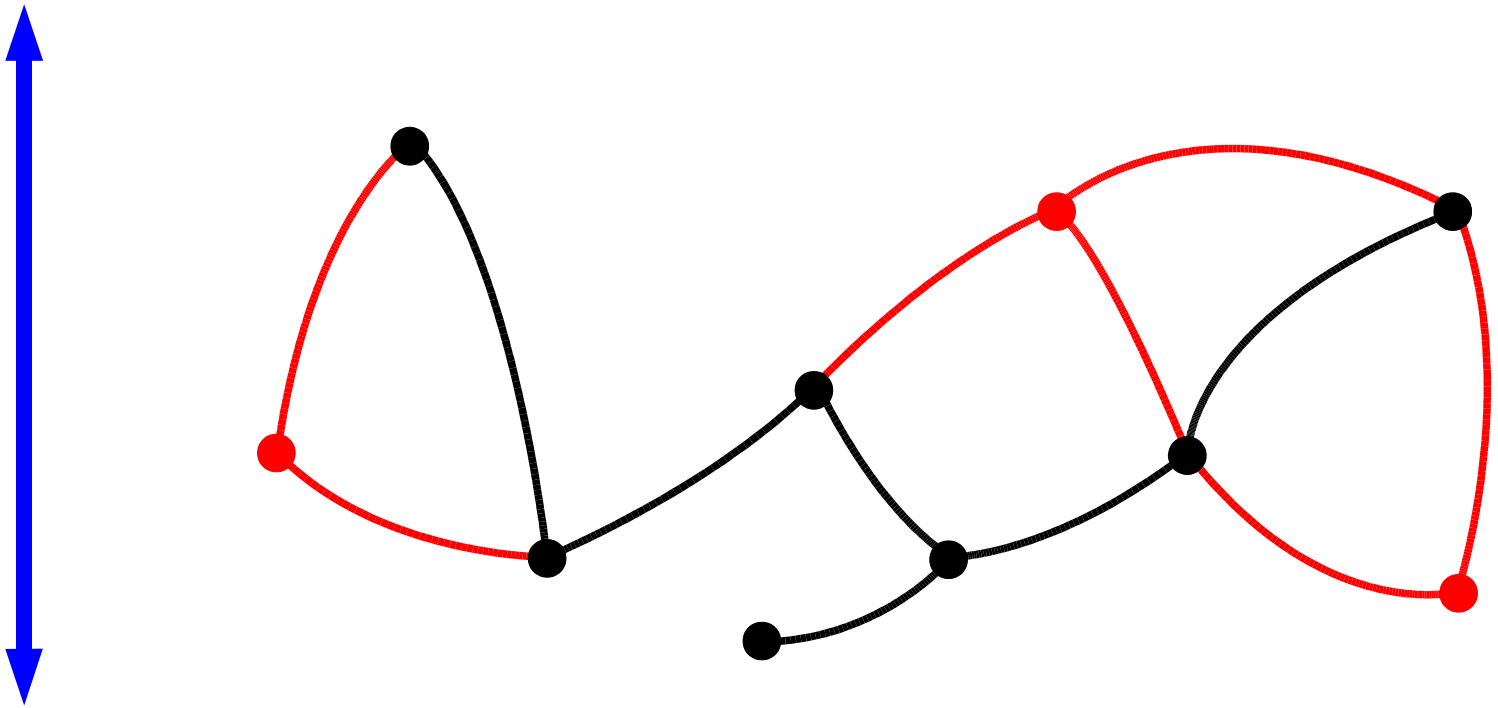


split of V: $a\ b \mid$ other vertices of G

Split system $\mathbf{S}_G$

The general problem is NP-hard

Splits that contribute to the PD-score of a subset.



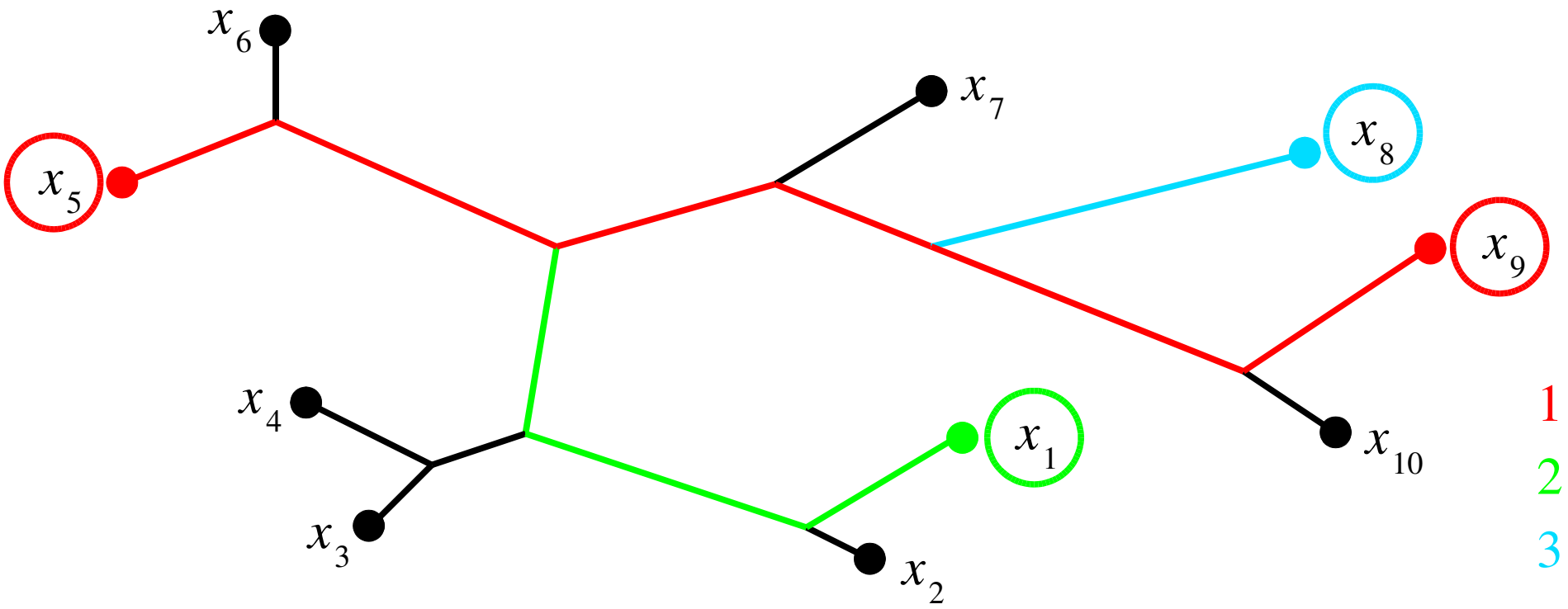
Edges in the graph “covered” by the subset.

Reduction from *Minimum Vertex Cover*.

Compatible split systems (phylogenetic trees)

Select elements greedily.

Steel, 2005
Pardi and Goldman, 2005

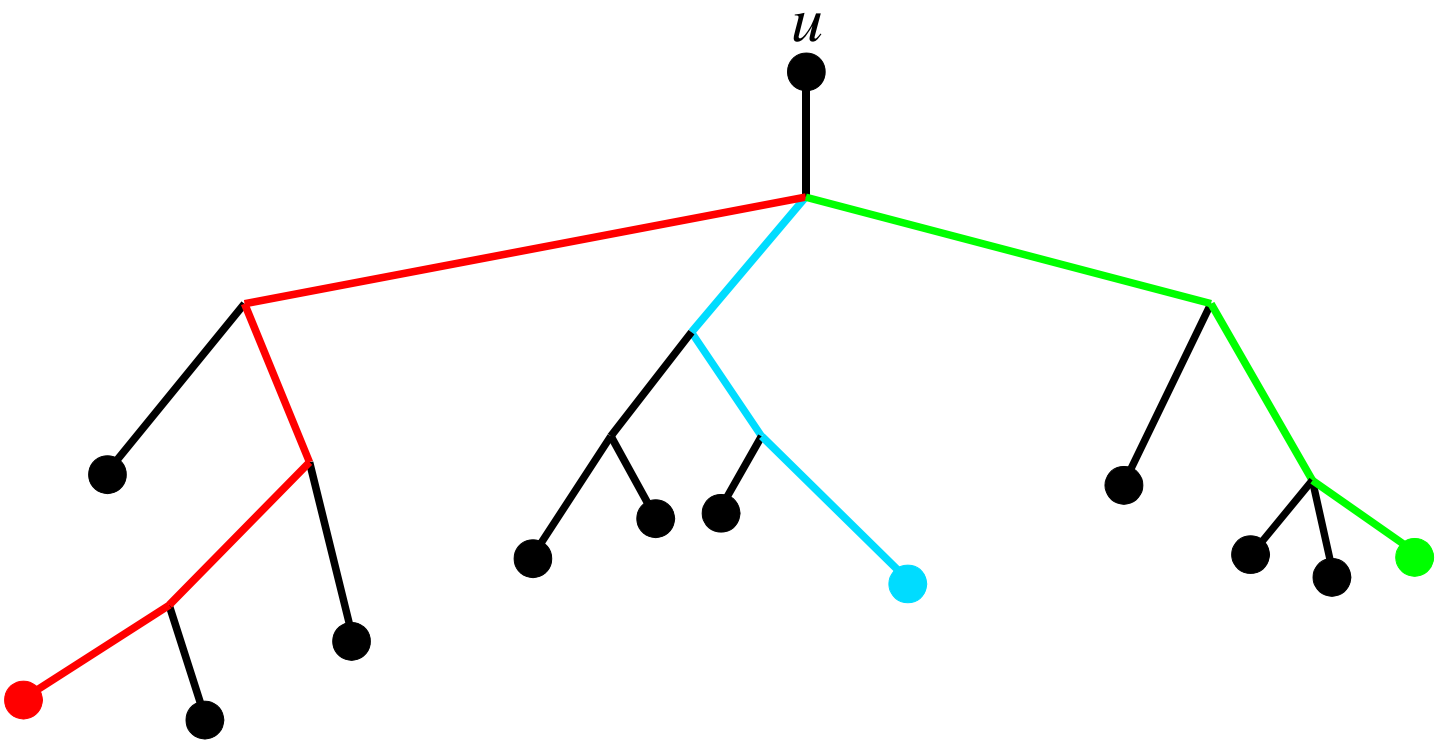


Implementation with run time $O(n \log k)$.

Minh et al., 2006

Compatible split systems (phylogenetic trees)

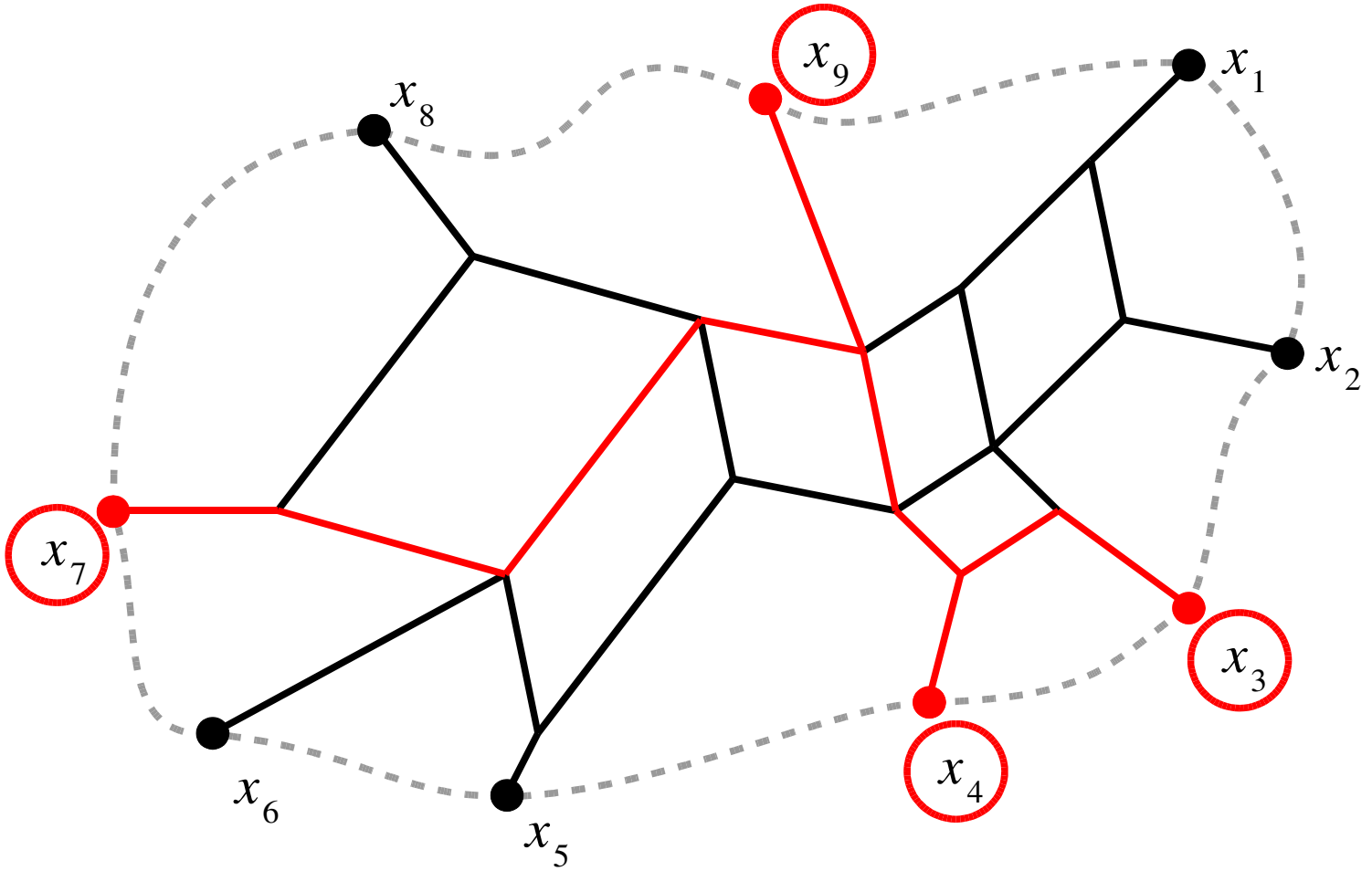
The longest path in the tree is an extension of the longest path in one of the subtrees



Implementation with run time $O(n)$.

Circular split systems

Respresentation by a network

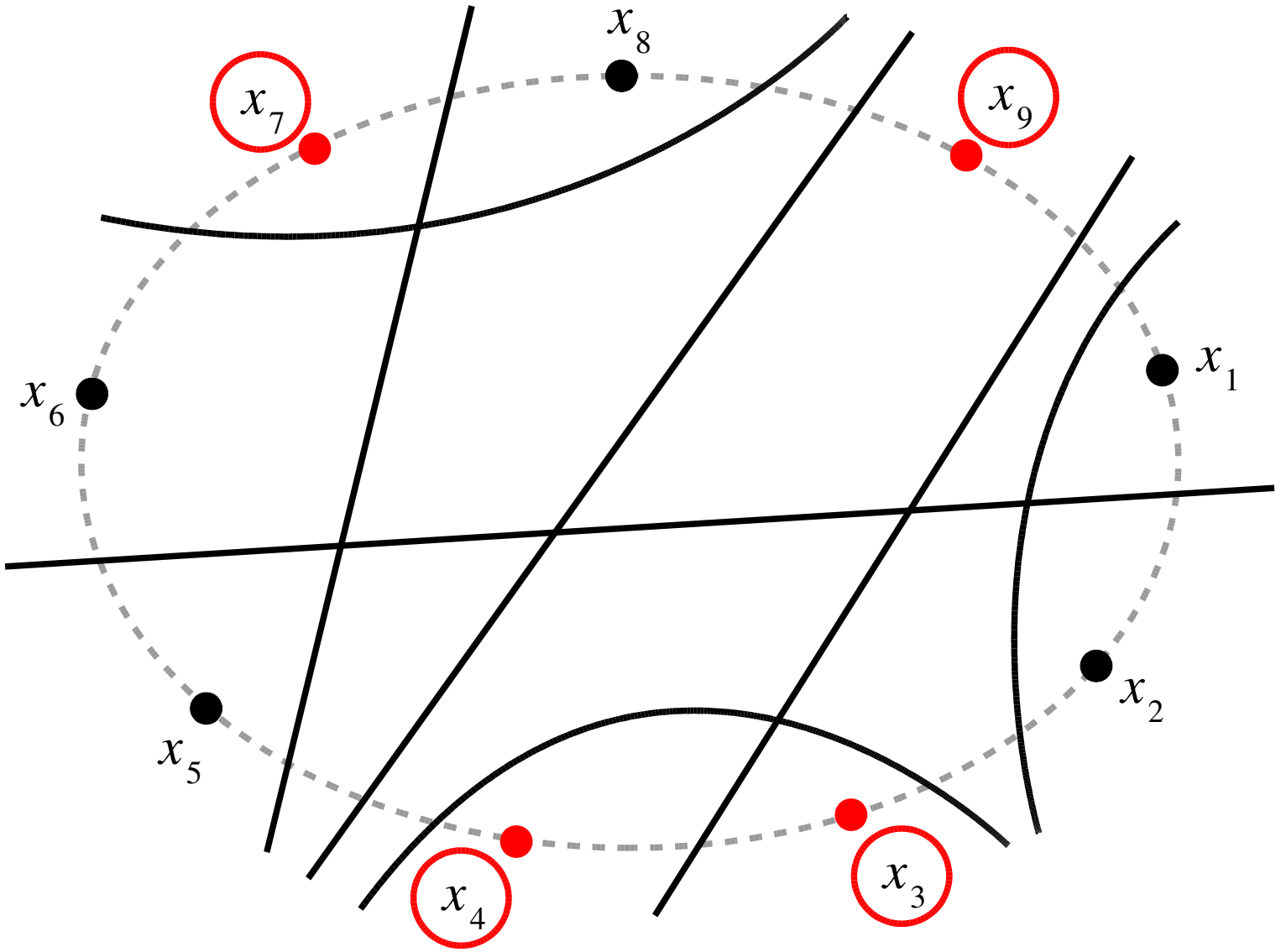


$O(kn^3)$ time algorithm based on dynamic programming

Minh et al., 2007

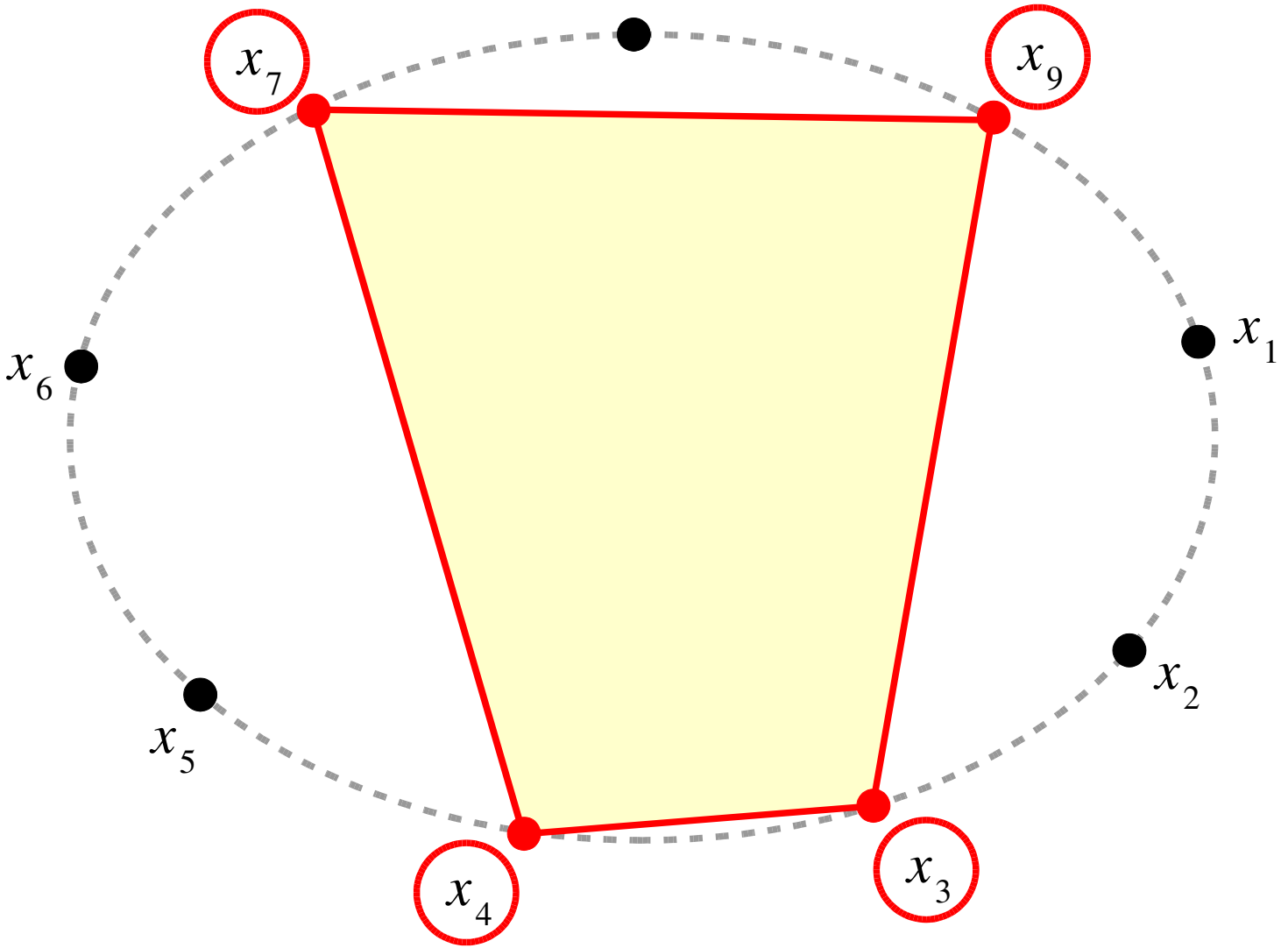
Circular split systems

Geometric representation by points and lines



Circular split systems

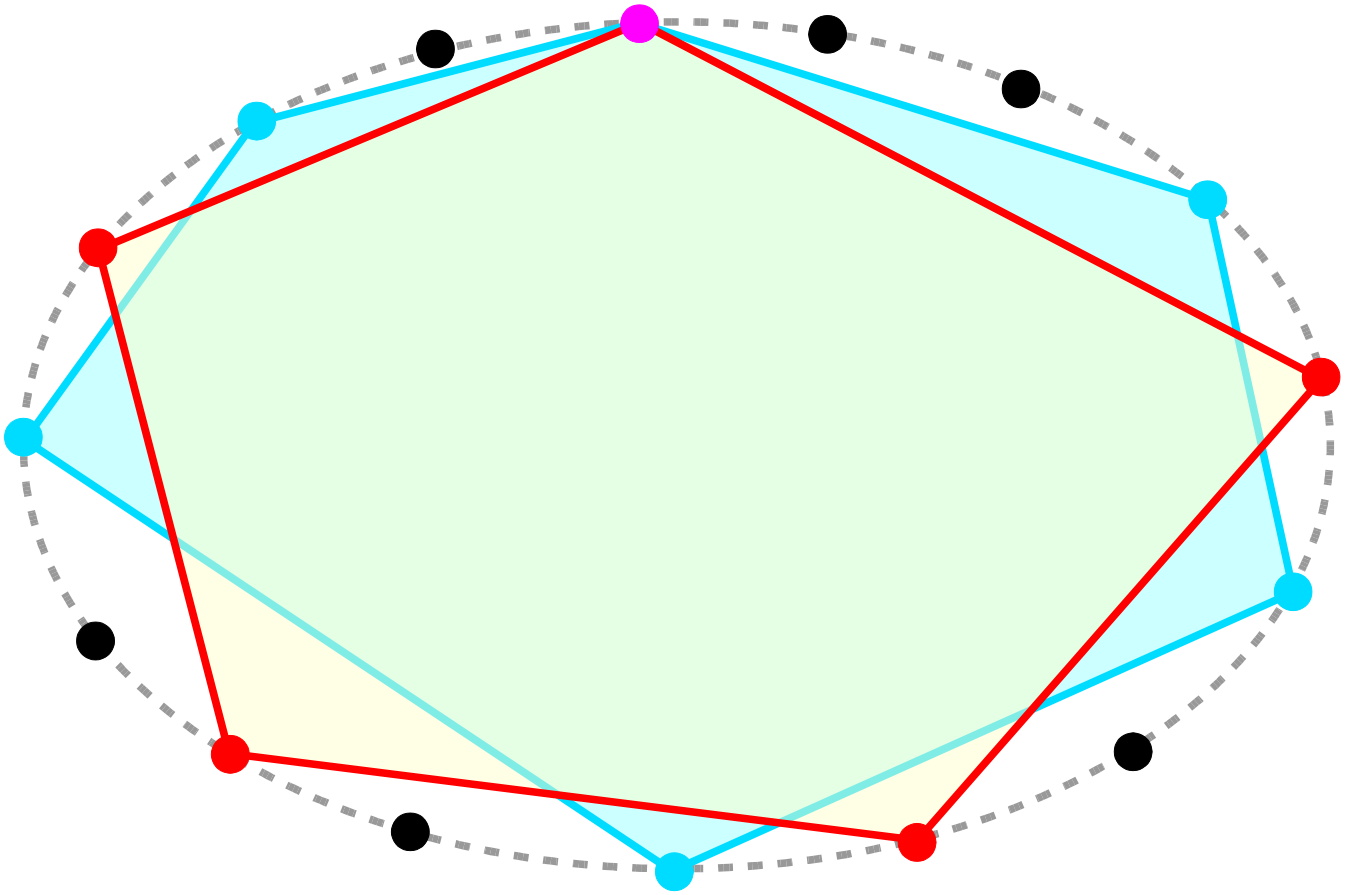
Find a convex polygon with k vertices and maximum perimeter.



Circular split systems

Key observation: optimal solutions interleave.

Boyce et al., 1985



Algorithm with run time $O(k\,n\ +\ n \log n)$.

Summary

- Explored structural properties of split systems that can help to solve the PD-optimization problem.
- Showed that the PD-optimization problem is hard even for some split systems of a very simple structure.
- Pointed out connections to another previously studied optimization problem.

T h a n k y o u