Nature Reserve Selection and Other Problems in Phylogenetic Diversity

Charles Semple

Allan Wilson Centre for Molecular Ecology and Evolution
&
Biomathematics Research Centre
Department of Mathematics and Statistics
University of Canterbury
New Zealand

Current Challenges and Problems in Phylogenetics
Isaac Newton Institute for Mathematical Sciences, 2007

Joint work with Magnus Bordewich

1. Motivation and Background

Conservation Biology

A central question in conservation biology is how to measure, predict, and preserve biodiversity as species face extinction.

Phylogenetic diversity (PD) is a quantative tool for measuring the biodiversity of a collection of species.

- PD is based on the evolutionary distance of the species.

(Faith 1992)

1. Motivation and Background

Conservation Biology

A central question in conservation biology is how to measure, predict, and preserve biodiversity as species face extinction.

Phylogenetic diversity (PD) is a quantative tool for measuring the biodiversity of a collection of species.

- PD is based on the evolutionary distance of the species.

(Faith 1992)

PD is one of many such measures
for selecting species.

1. Motivation and Background

Conservation Biology

A central question in conservation biology is how to measure, predict, and preserve biodiversity as species face extinction.

Phylogenetic diversity (PD) is a quantative tool for measuring the biodiversity of a collection of species.

- PD is based on the evolutionary distance of the species.

(Faith 1992)

PD is one of many such measures for selecting species.

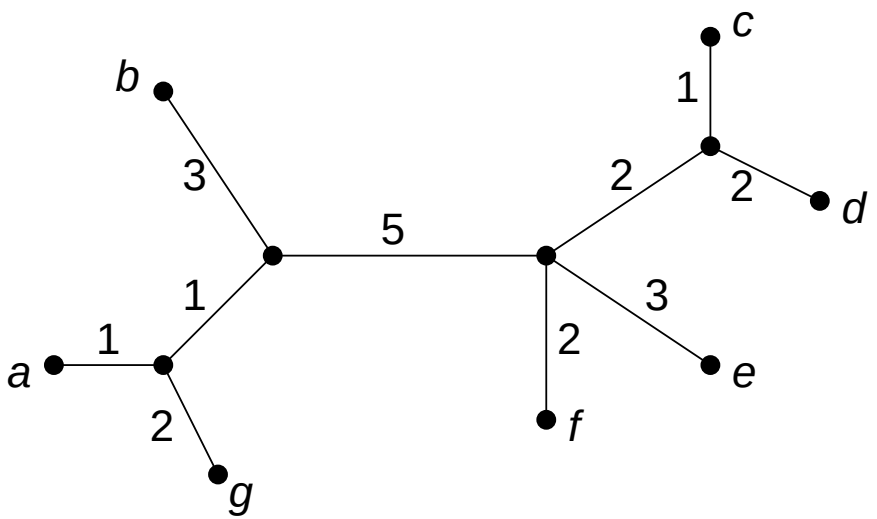
. . . it would be hard to justify ruling against Sphenodon (tuatara) under any scheme.

(Crozier 1997)

Phylogenetic Diversity

For an edge-weighted phylogenetic X-tree T, the phylogenetic diversity (PD) score of a subset $S \subseteq X$ is the sum of the weights of the edges of T spanned by S.

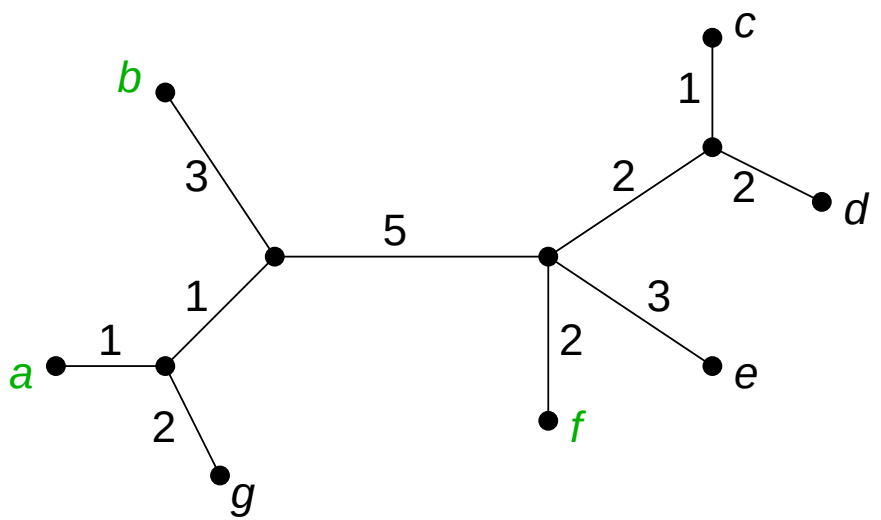
Example



Phylogenetic Diversity

For an edge-weighted phylogenetic X-tree T, the phylogenetic diversity (PD) score of a subset $S \subseteq X$ is the sum of the weights of the edges of T spanned by S.

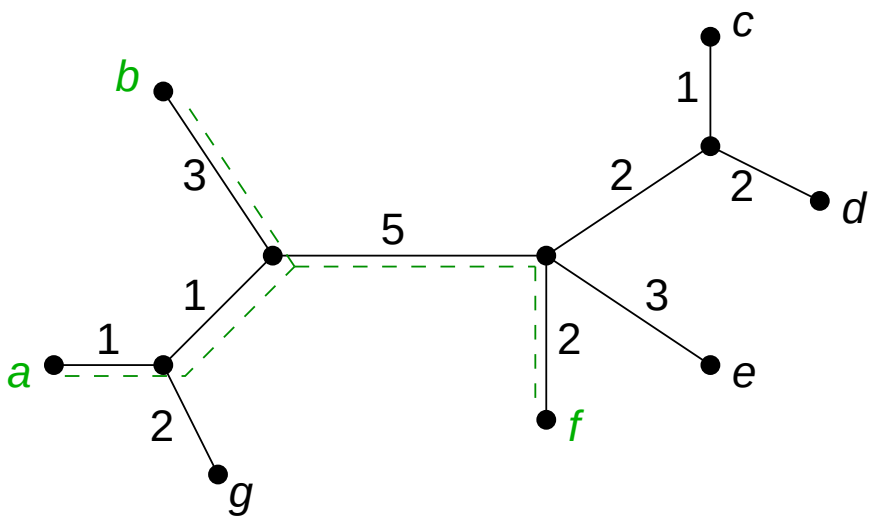
Example



Phylogenetic Diversity

For an edge-weighted phylogenetic X-tree T, the phylogenetic diversity (PD) score of a subset $S \subseteq X$ is the sum of the weights of the edges of T spanned by S.

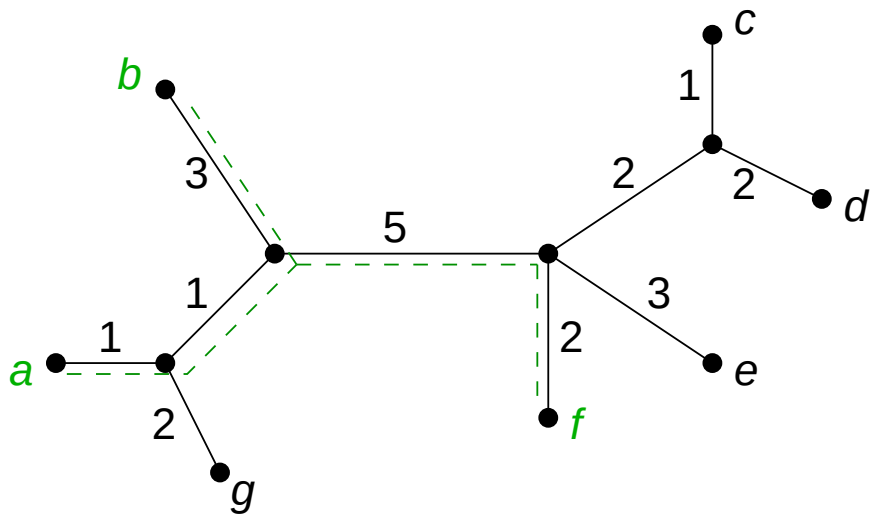
Example



Phylogenetic Diversity

For an edge-weighted phylogenetic X-tree T, the phylogenetic diversity (PD) score of a subset $S \subseteq X$ is the sum of the weights of the edges of T spanned by S.

Example



$PD(\{a, b, f\}) = 12$

Basic Problem

Given

- an edge-weighted phylogenetic X-tree T
- a positive integer $k \geq 2$

Find a subset of X of size k that maximizes the PD score on T over all such subsets.

Basic Problem

Given

- an edge-weighted phylogenetic X-tree T
- a positive integer $k \geq 2$

Find a subset of X of size k that maximizes the PD score on T over all such subsets.

- This optimization problem is the **Basic Problem**.

Basic Problem

Given

- an edge-weighted phylogenetic X-tree T
- a positive integer $k \geq 2$

Find a subset of X of size k that maximizes the PD score on T over all such subsets.

- This optimization problem is the Basic Problem.

Theorem (Pardi, Goldmann 2005; Steel 2005)

Basic Problem is solvable in polynomial time.

2. Nature Reserve Selection Problem

Conserving Whole Habitats Instead of Single Species

In practice, one frequently does not conserve species in isolation and one has to work within a budget.

2. Nature Reserve Selection Problem

Conserving Whole Habitats Instead of Single Species

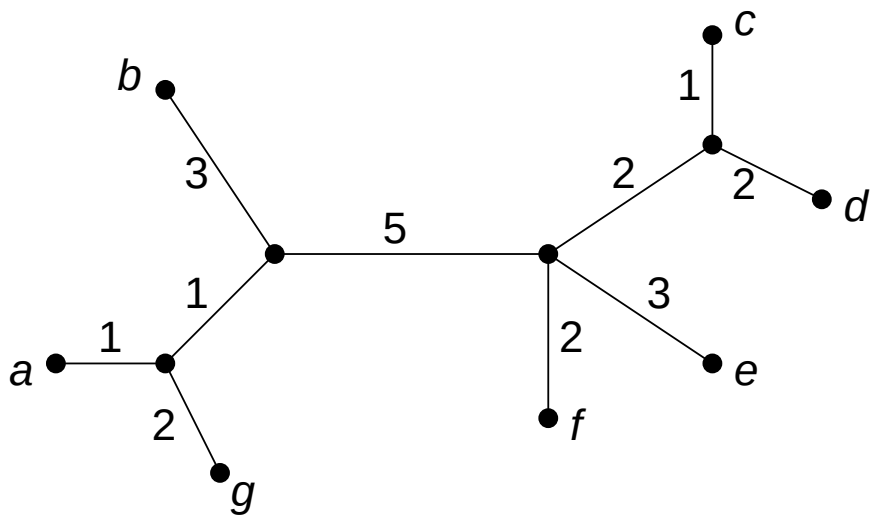
In practice, one frequently does not conserve species in isolation and one has to work within a budget.

> *Although conservation action is frequently targeted towards single species, the most effective way of preserving overall species diversity is by conserving viable populations in their natural habitats, often by designating networks of protected areas.*
>
> (Rodrigues, Brooks, Gaston 2005)

Extending Phylogenetic Diversity to Areas

For an edge-weighted phylogenetic X-tree T, and a collection $\mathcal{A}$ of protected areas containing species in X, the phylogenetic diversity (PD) score of a subset $S \subseteq \mathcal{A}$ is the PD score of the species contained within at least one area in S.

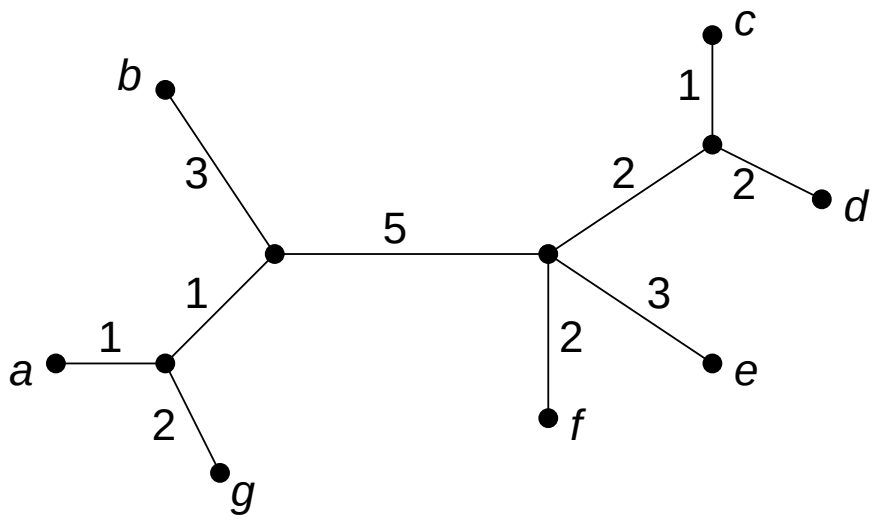
Example



Extending Phylogenetic Diversity to Areas

For an edge-weighted phylogenetic X-tree T, and a collection $\mathcal{A}$ of protected areas containing species in X, the phylogenetic diversity (PD) score of a subset $S \subseteq \mathcal{A}$ is the PD score of the species contained within at least one area in S.

Example

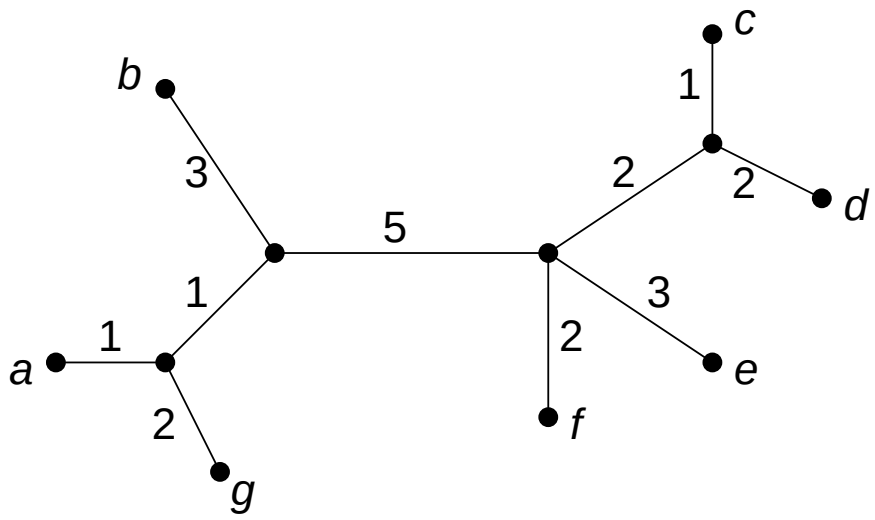


$S = \{\{a, b\}, \{c, e\}, \{a, g, e\}\}$

Extending Phylogenetic Diversity to Areas

For an edge-weighted phylogenetic X-tree T, and a collection $\mathcal{A}$ of protected areas containing species in X, the phylogenetic diversity (PD) score of a subset $S \subseteq \mathcal{A}$ is the PD score of the species contained within at least one area in S.

Example



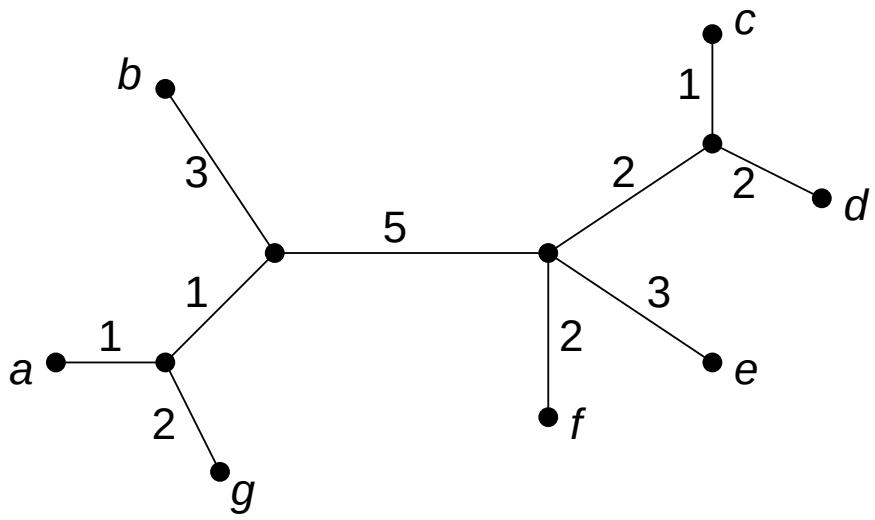
$S = \{\{a, b\}, \{c, e\}, \{a, g, e\}\}$

$PD(\{\{a, b\}, \{c, e\}, \{a, g, e\}\}) =$

Extending Phylogenetic Diversity to Areas

For an edge-weighted phylogenetic X-tree T, and a collection $\mathcal{A}$ of protected areas containing species in X, the phylogenetic diversity (PD) score of a subset $S \subseteq \mathcal{A}$ is the PD score of the species contained within at least one area in S.

Example



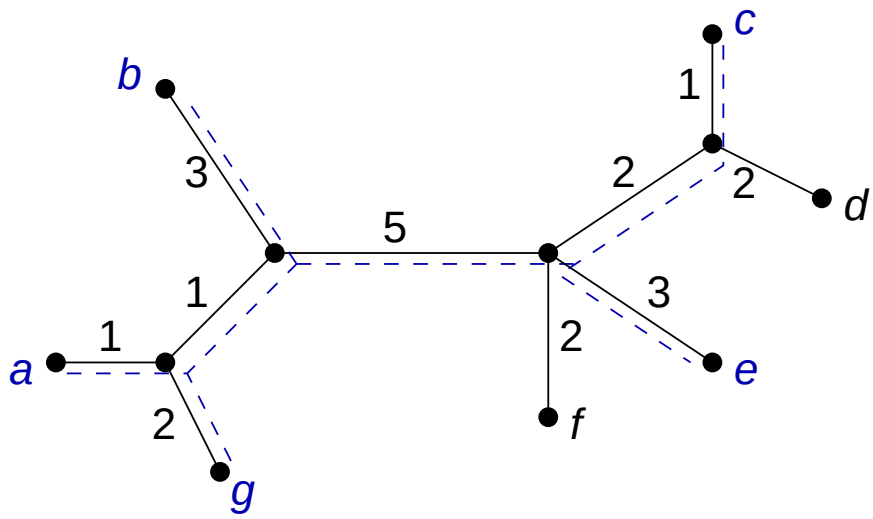
$S = \{\{a, b\}, \{c, e\}, \{a, g, e\}\}$

$PD(\{\{a, b\}, \{c, e\}, \{a, g, e\}\}) = PD(\{a, b, c, e, g\}) =$

Extending Phylogenetic Diversity to Areas

For an edge-weighted phylogenetic X-tree T, and a collection $\mathcal{A}$ of protected areas containing species in X, the phylogenetic diversity (PD) score of a subset $S \subseteq \mathcal{A}$ is the PD score of the species contained within at least one area in S.

Example



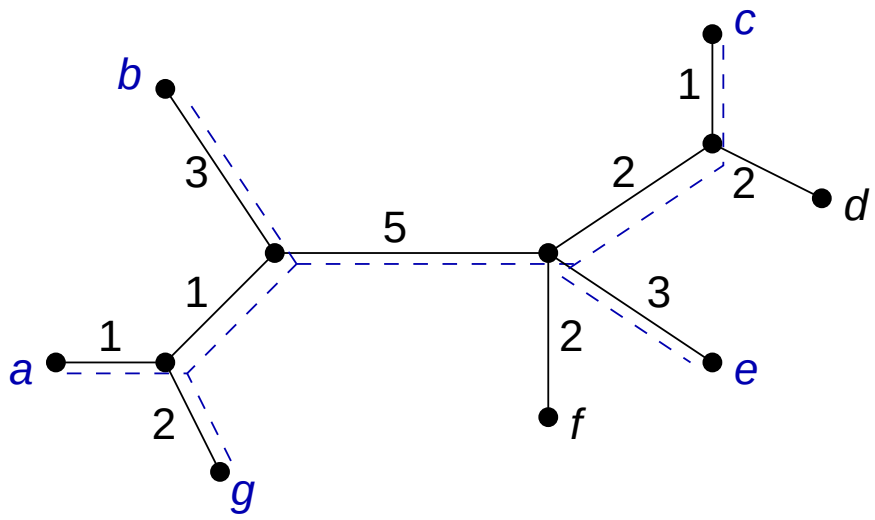
$S = \{\{a, b\}, \{c, e\}, \{a, g, e\}\}$

$PD(\{\{a, b\}, \{c, e\}, \{a, g, e\}\}) = PD(\{a, b, c, e, g\}) =$

Extending Phylogenetic Diversity to Areas

For an edge-weighted phylogenetic X-tree T, and a collection $\mathcal{A}$ of protected areas containing species in X, the phylogenetic diversity (PD) score of a subset $S \subseteq \mathcal{A}$ is the PD score of the species contained within at least one area in S.

Example



$S = \{\{a, b\}, \{c, e\}, \{a, g, e\}\}$

$PD(\{\{a, b\}, \{c, e\}, \{a, g, e\}\}) = PD(\{a, b, c, e, g\}) = 18$

Nature Reserve Selection Problem

GIVEN

- an edge-weight phylogenetic X-tree T
- a collection $\mathcal{A}$ of (conservation) areas containing species in X
- a preservation cost for each area
- a fixed budget B

FIND a subset of areas in $\mathcal{A}$ to preserve that maximizes the PD score on T of the species contained within the preserved areas while keeping within B.

This problem is called the BUDGETED NATURE RESERVE SELECTION problem (BRNS).

Nature Reserve Selection Problem

GIVEN

- an edge-weight phylogenetic X-tree T
- a collection $\mathcal{A}$ of (conservation) areas containing species in X
- a preservation cost for each area
- a fixed budget B

FIND a subset of areas in $\mathcal{A}$ to preserve that maximizes the PD score on T of the species contained within the preserved areas while keeping within B.

This problem is called the BUDGETED NATURE RESERVE SELECTION problem (BRNS).

For applications of BNRS in conservation planning, see Moritz and Faith 1998, Rodrigues and Gaston 2002, Smith et al 2000.

BNRS

Theorem (Moulton, S, Steel 2007)

Even if each area in $\mathcal{A}$ has unit cost, BNRS is NP-hard.

BNRS

Theorem (Moulton, S, Steel 2007)

Even if each area in $\mathcal{A}$ has unit cost, BNRS is NP-hard.

Theorem (Bordewich, S 2007)

There is a polynomial-time $(1 - 1/e)$-approximation algorithm for BNRS.

BNRS

Theorem (Moulton, S, Steel 2007)

Even if each area in $\mathcal{A}$ has unit cost, BNRS is NP-hard.

Theorem (Bordewich, S 2007)

There is a polynomial-time $(1 - 1/e)$-approximation algorithm for BNRS.

- $(1 - 1/e) \approx 63\%$

BNRS

Theorem (Moulton, S, Steel 2007)

Even if each area in $\mathcal{A}$ has unit cost, BNRS is NP-hard.

Theorem (Bordewich, S 2007)

There is a polynomial-time $(1 - 1/e)$-approximation algorithm for BNRS. Moreover, for any $\epsilon > 0$, BNRS cannot be approximated with an approximation ratio of $(1 - 1/e + \epsilon)$ unless P=NP.

- $(1 - 1/e) \approx 63\%$

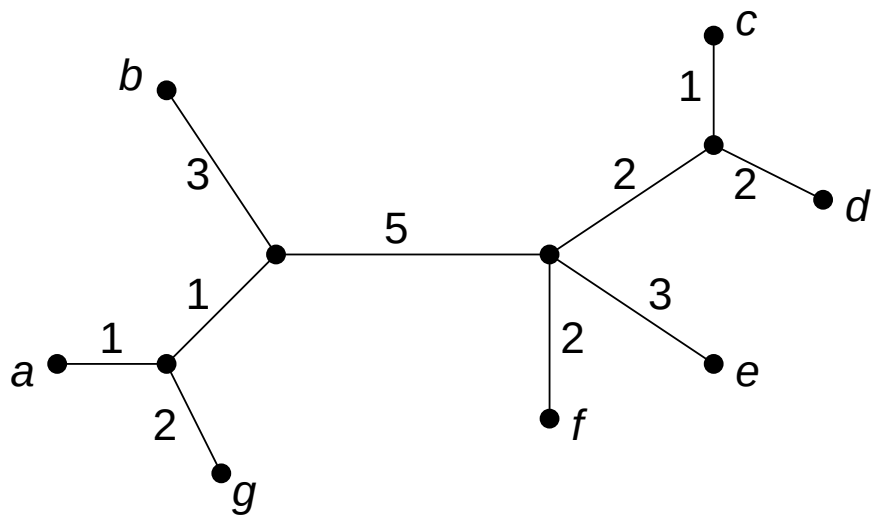
Approximation Solution to BNRS

1. Exhaustively find a feasible solution of size at most 2 that maximizes the PD score. Call the resulting solution H_1.
2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.
3. Call the best solution in 2. H_2.
4. Output H_1 or H_2 depending on which has the bigger PD.

Approximation Solution to BNRS

2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.

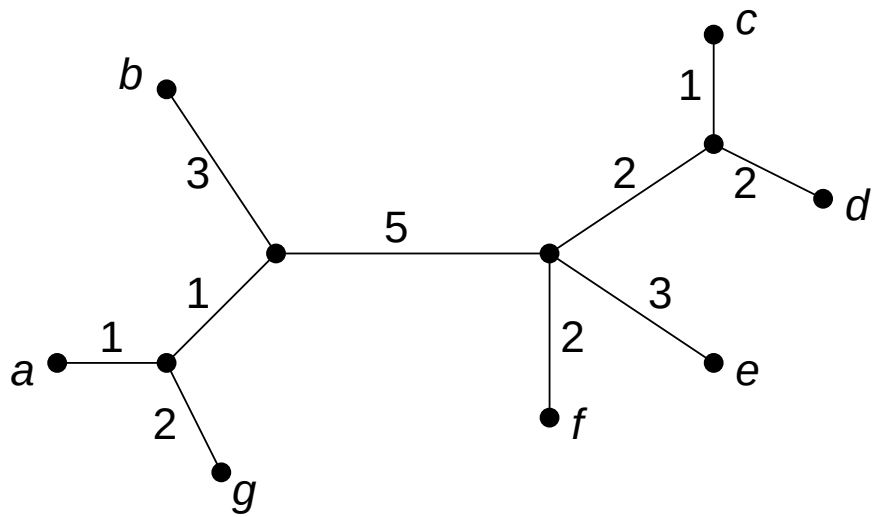
Example. $\mathcal{A} = \{\{b\}, \{f, c\}, \{c, d\}, \{a, b\}, \{a, g\}, \{e\}, \{g, e\}\}$ and $B = 24$



Approximation Solution to BNRS

2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.

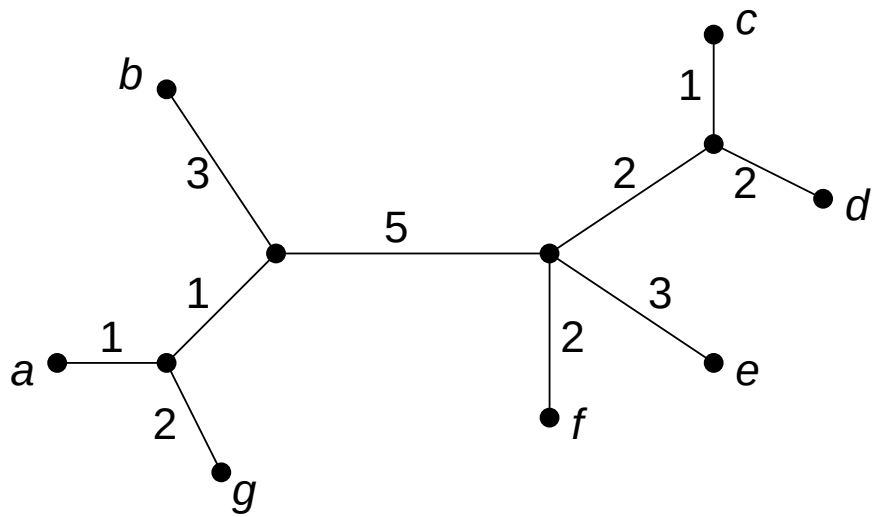
Example. $c(\{b\}) = 2$, $c(\{f, c\}) = 8$, $c(\{c, d\}) = 6$, $c(\{a, b\}) = 10$, $c(\{a, g\}) = 4$, $c(\{e\}) = 4$, $c(\{g, e\}) = 5$ and $B = 24$



Approximation Solution to BNRS

2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.

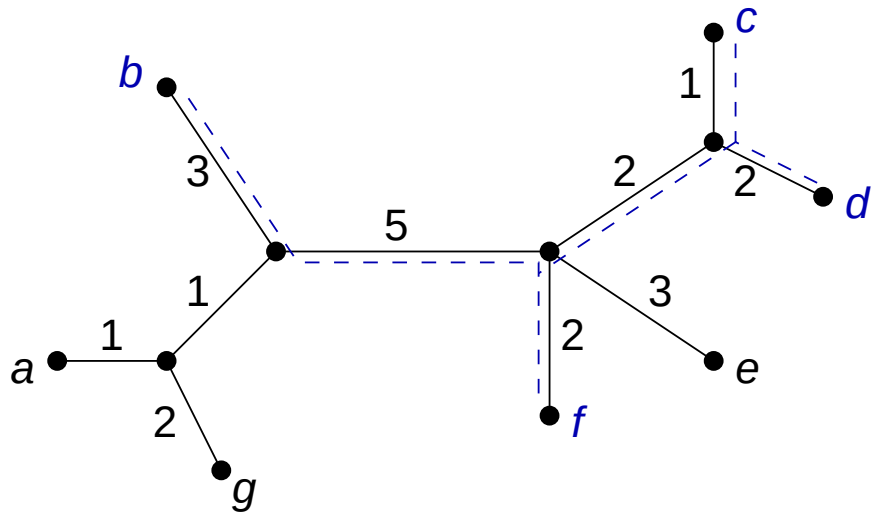
Example. $c(\{b\}) = 2$, $c(\{f, c\}) = 8$, $c(\{c, d\}) = 6$, $c(\{a, b\}) = 10$, $c(\{a, g\}) = 4$, $c(\{e\}) = 4$, $c(\{g, e\}) = 5$ and $B = 24$



Approximation Solution to BNRS

2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.

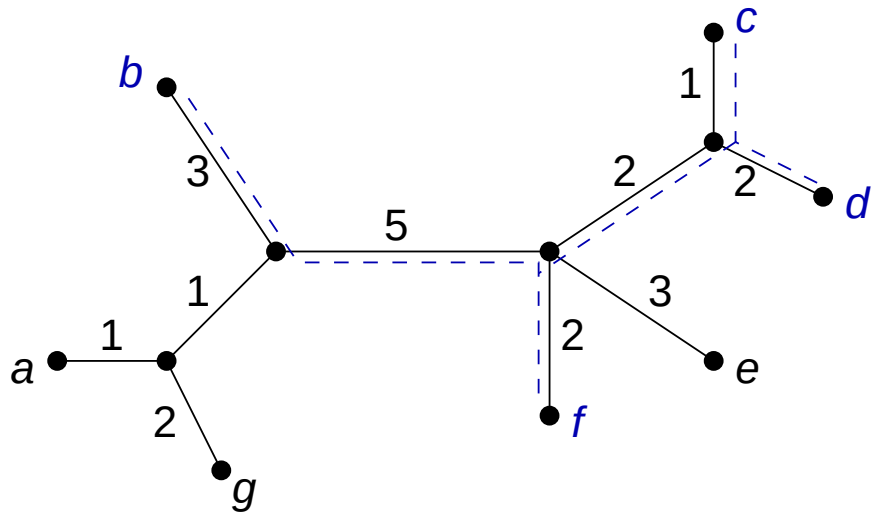
Example. $c(\{b\}) = 2$, $c(\{f, c\}) = 8$, $c(\{c, d\}) = 6$, $c(\{a, b\}) = 10$, $c(\{a, g\}) = 4$, $c(\{e\}) = 4$, $c(\{g, e\}) = 5$ and $B = 24$



Approximation Solution to BNRS

2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.

Example. $c(\{b\}) = 2$, $c(\{f, c\}) = 8$, $c(\{c, d\}) = 6$, $c(\{a, b\}) = 10$, $c(\{a, g\}) = 4$, $c(\{e\}) = 4$, $c(\{g, e\}) = 5$ and $B = 24$

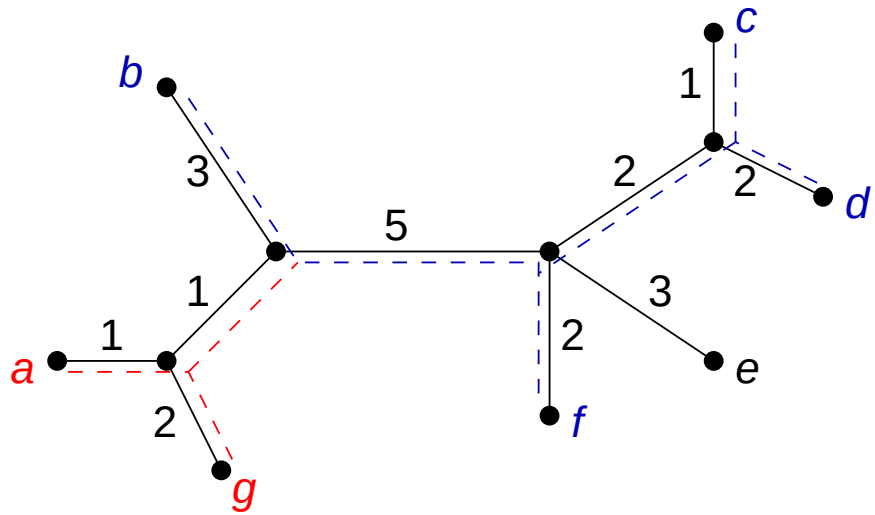


Approximation Solution to BNRS

2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.

Example. $c(\{b\}) = 2$, $c(\{f,c\}) = 8$, $c(\{c,d\}) = 6$, $c(\{a,b\}) = 10$, $c(\{a,g\}) = 4$, $c(\{e\}) = 4$, $c(\{g,e\}) = 5$ and $B = 24$

$$r = \frac{PD(B \cup \{a,g\}) - PD(B)}{c(\{a,g\})} = 4/4$$



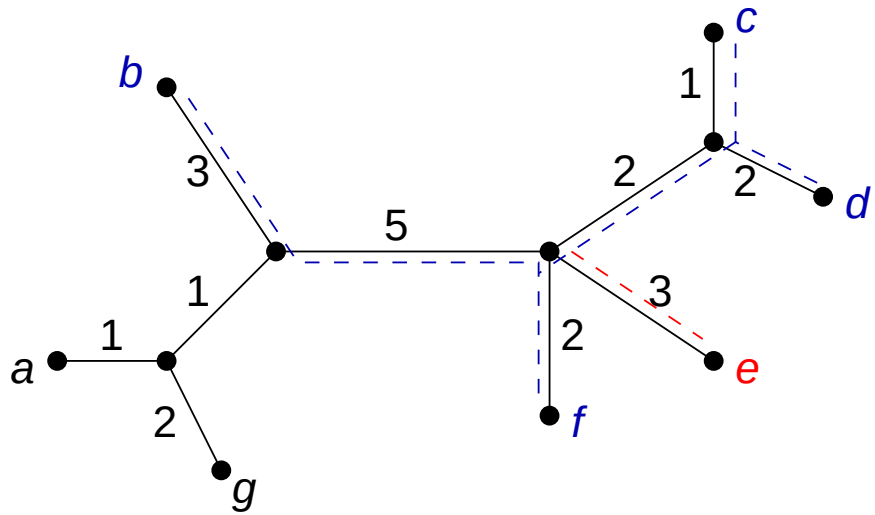
Approximation Solution to BNRS

2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.

Example. $c(\{b\}) = 2$, $c(\{f,c\}) = 8$, $c(\{c,d\}) = 6$, $c(\{a,b\}) = 10$, $c(\{a,g\}) = 4$, $c(\{e\}) = 4$, $c(\{g,e\}) = 5$ and $B = 24$

$$r = \frac{PD(B \cup \{a,g\}) - PD(B)}{c(\{a,g\})} = 4/4$$

$$r = \frac{PD(B \cup \{e\}) - PD(B)}{c(\{e\})} = 3/4$$



Approximation Solution to BNRS

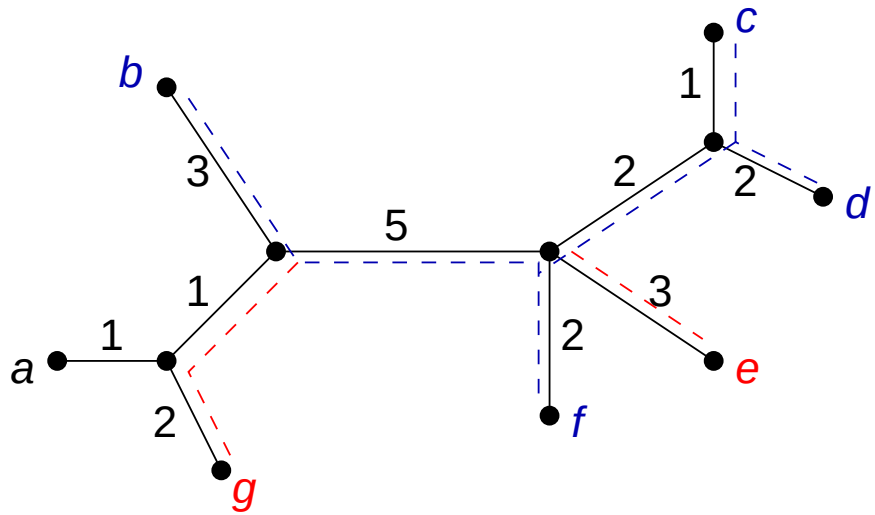
2. For all subsets of $\mathcal{A}$ of size 3,
   (a) Sequentially add areas from $\mathcal{A}$ that maximize the ratio of incremental PD to cost while keeping within budget.
   (b) Do this until no more areas can be added.

Example. $c(\{b\}) = 2$, $c(\{f,c\}) = 8$, $c(\{c,d\}) = 6$, $c(\{a,b\}) = 10$, $c(\{a,g\}) = 4$, $c(\{e\}) = 4$, $c(\{g,e\}) = 5$ and $B = 24$

$$r = \frac{PD(B \cup \{a,g\}) - PD(B)}{c(\{a,g\})} = 4/4$$

$$r = \frac{PD(B \cup \{e\}) - PD(B)}{c(\{e\})} = 3/4$$

$$r = \frac{PD(B \cup \{g,e\}) - PD(B)}{c(\{g,e\})} = 6/5$$



Remarks

- BNRS extends the problems BUDGETED MAXIMUM COVERAGE and MAXIMUM k-COVERAGE.
  - Khuller, Moss, Naor 1999 showed that there is a polynomial-time $(1 - 1/e)$-approximation algorithm for BUDGETED MAXIMUM COVERAGE.
  - Feige 1998 showed that, for any $\epsilon > 0$, MAXIMUM k-COVERAGE cannot be approximated with a approximation ratio of $(1 - 1/e + \epsilon)$ unless P=NP

Remarks

- If all conservation areas have the same preservation cost, then the following greedy approach achieves the approximation ratio $(1 - 1/e) \approx 63\%$.

  1. Exhaustively find a subset H_0 of size 2 that maximizes the PD score over all such subsets.
  2. Sequentially add areas from $\mathcal{A}$ to H_0 that maximize the incremental increase in PD while keeping within budget.
  3. Do this until no more areas can be added. Output the resulting solution.

Remarks

- If all conservation areas have the same preservation cost, then the following greedy approach achieves the approximation ratio $(1 - 1/e) \approx 63\%$.

  1. Exhaustively find a subset H_0 of size 2 that maximizes the PD score over all such subsets.
  2. Sequentially add areas from $\mathcal{A}$ to H_0 that maximize the incremental increase in PD while keeping within budget.
  3. Do this until no more areas can be added. Output the resulting solution.

  Because of Feige's result, this approximation ratio is the best possible (unless P=NP).

3. Other Problems

Optimizing Diversity with Coverage

Given

- an edge-weighted phylogenetic X-tree T
- a collection $\mathcal{A}$ of subsets of X with each subset identifying the species with a particular characteristic
- a threshold n_A for each characteristic in $\mathcal{A}$
- a positive integer k

Find a subset of X of size at most k that maximizes the PD score on T such that for each characteristic its threshold is satisfied.

This problem is called Optimizing Diversity with Coverage.

Some good news

If the characteristics are pairwise disjoint and the subtrees in $\{T(A) : A \in \mathcal{A}\}$ are vertex disjoint, then the problem is solvable in polynomial-time. (Moulton, S, Steel 2007)

Optimizing Diversity with Coverage

Some good news

If the characteristics are pairwise disjoint and the subtrees in $\{T(A) : A \in \mathcal{A}\}$ are vertex disjoint, then the problem is solvable in polynomial-time. (Moulton, S, Steel 2007)

Plenty of bad news

Deciding if Optimizing Diversity with Coverage has a feasible solution is equivalent to Hitting Set. (Bordewich, S 2007)

Some good news

If the characteristics are pairwise disjoint and the subtrees in $\{T(A) : A \in \mathcal{A}\}$ are vertex disjoint, then the problem is solvable in polynomial-time. (Moulton, S, Steel 2007)

Plenty of bad news

Deciding if OPTIMIZING DIVERSITY WITH COVERAGE has a feasible solution is equivalent to HITTING SET. (Bordewich, S 2007)

Problem: HITTING SET
Instance: A collection $\mathcal{A}$ of subsets of X and a positive integer k.
Question: Is there a subset of X' of X of size k such that $A \cap X' \neq \emptyset$ for all $A \in \mathcal{A}$?

Some good news

If the characteristics are pairwise disjoint and the subtrees in $\{T(A) : A \in \mathcal{A}\}$ are vertex disjoint, then the problem is solvable in polynomial-time. (Moulton, S, Steel 2007)

Plenty of bad news

Deciding if OPTIMIZING DIVERSITY WITH COVERAGE has a feasible solution is equivalent to HITTING SET. (Bordewich, S 2007)

Problem: HITTING SET
Instance: A collection $\mathcal{A}$ of subsets of X and a positive integer k.
Question: Is there a subset of X' of X of size k such that $A \cap X' \neq \emptyset$ for all $A \in \mathcal{A}$?

This equivalence suggests that fairly severe restrictions are required to make OPTIMIZING DIVERSITY WITH COVERAGE solvable or even approximable.