Minimal Triangulation Algorithms for Perfect Phylogeny Problems

Rob Gysel

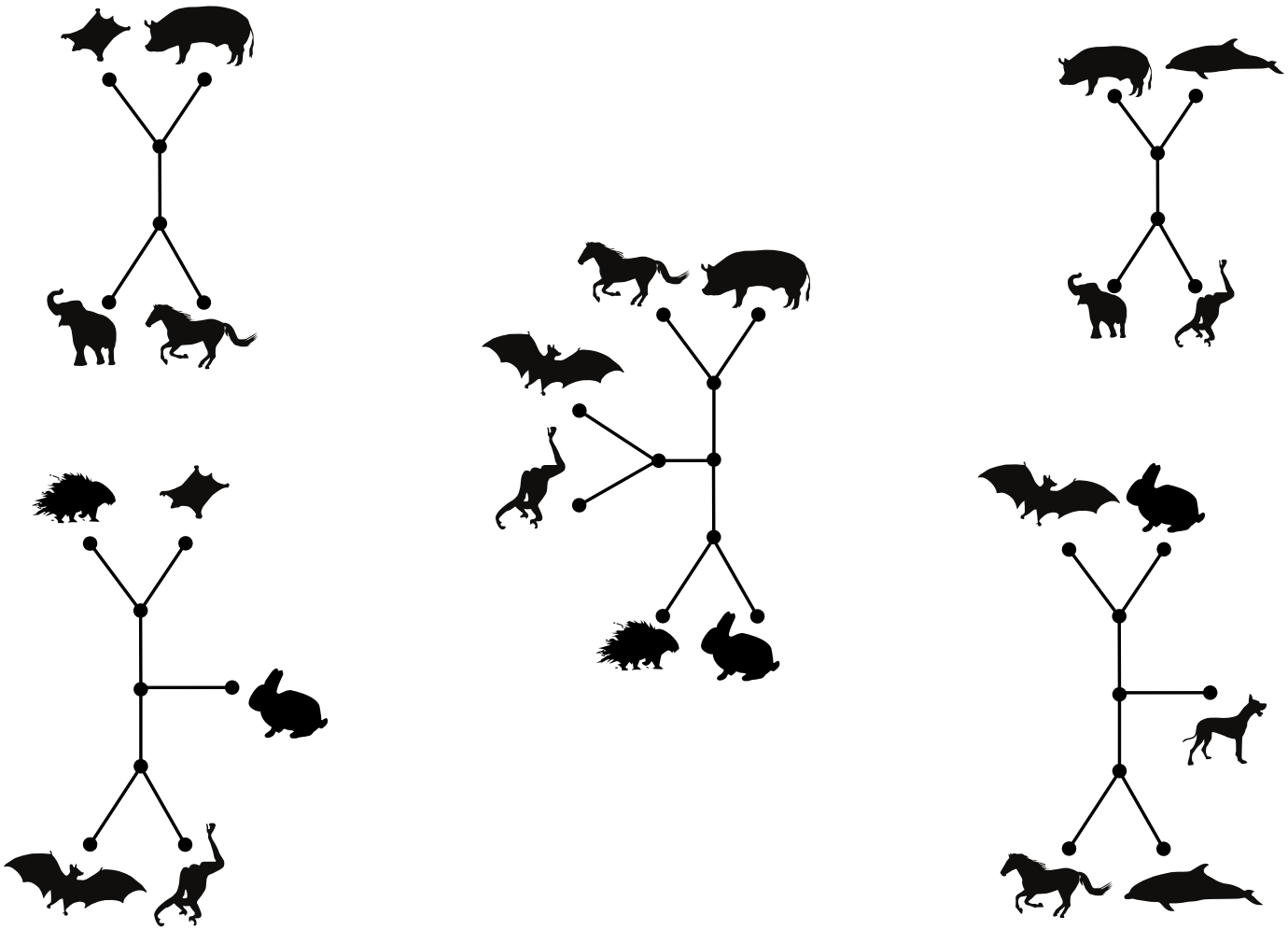
Department of Computer Science
University of California, Davis
2063 Kemper Hall, 1 Shields Avenue, Davis CA 95616

8th International Conference on Language and Automata Theory and Applications
March 11th, 2014

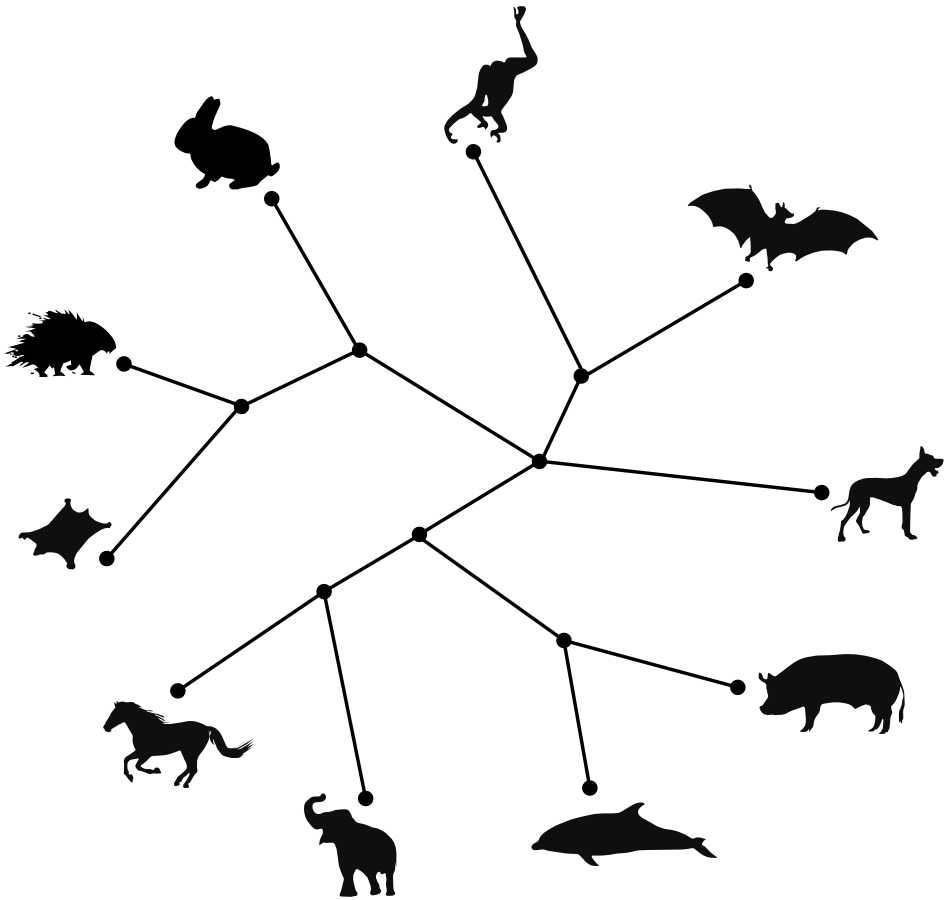
Overview

1. Motivation
2. Perfect Phylogeny / Character Compatibility Problems
3. Triangulation Framework
4. Minimal Triangulation Algorithms

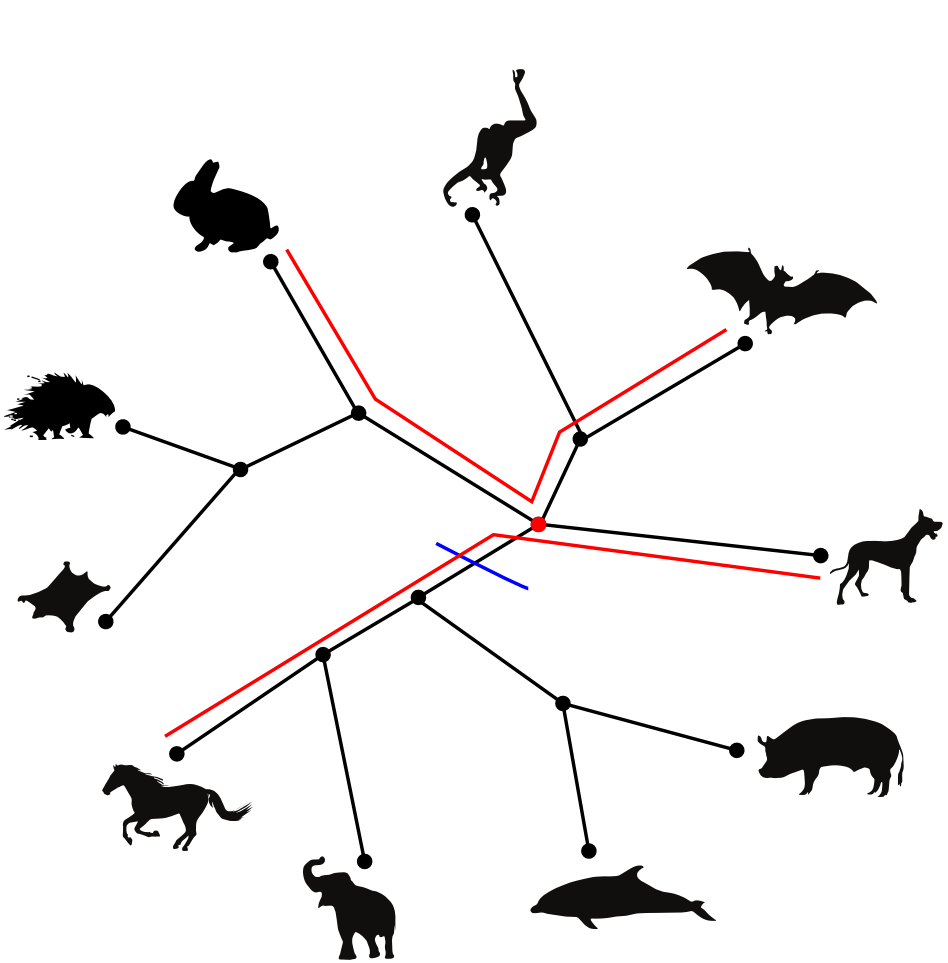
Motivation: Supertree Estimation

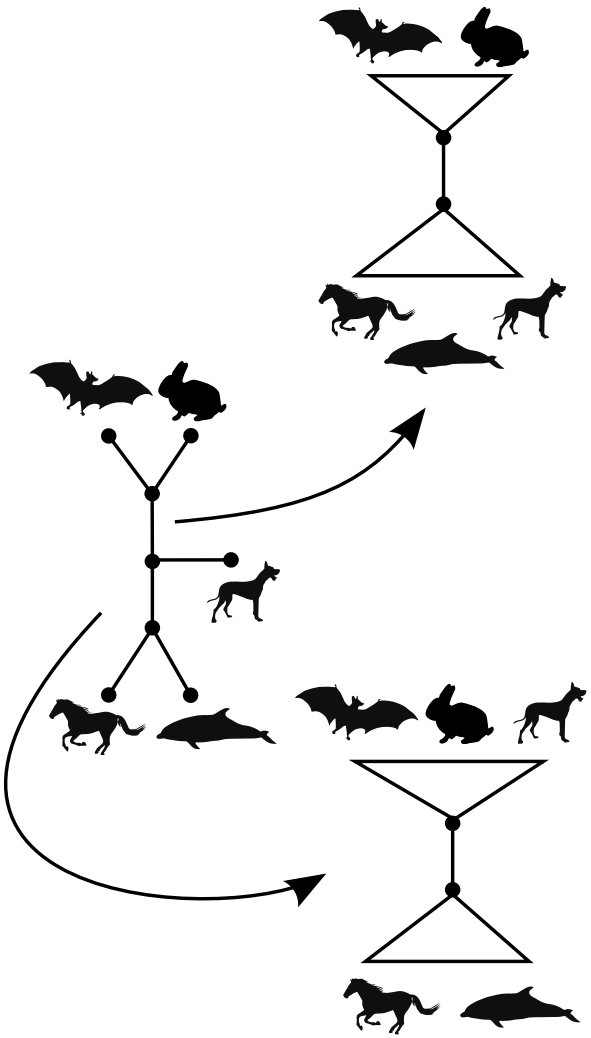


Motivation: Supertree Estimation

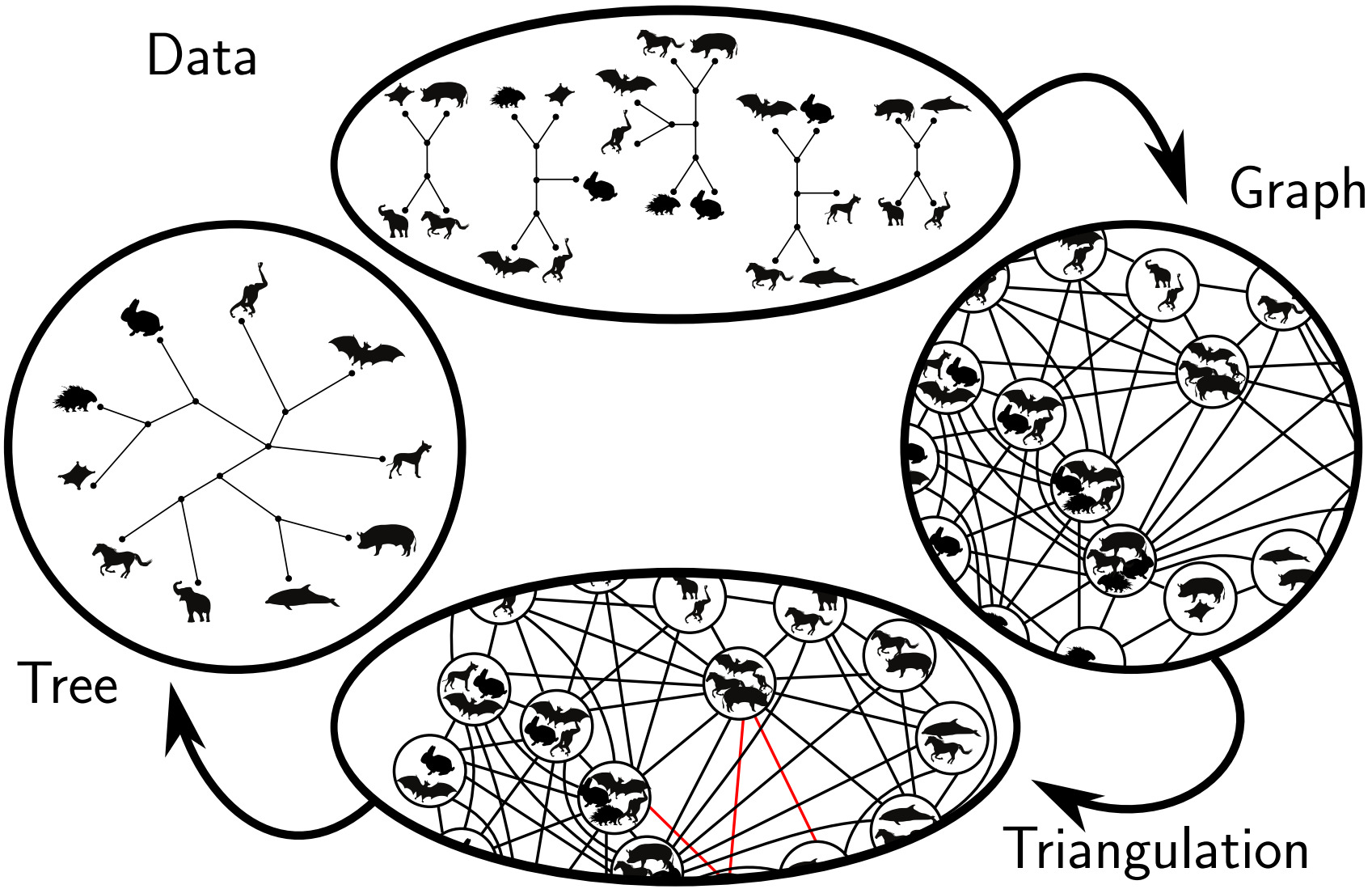


Motivation: Supertree Estimation





Triangulation Approach

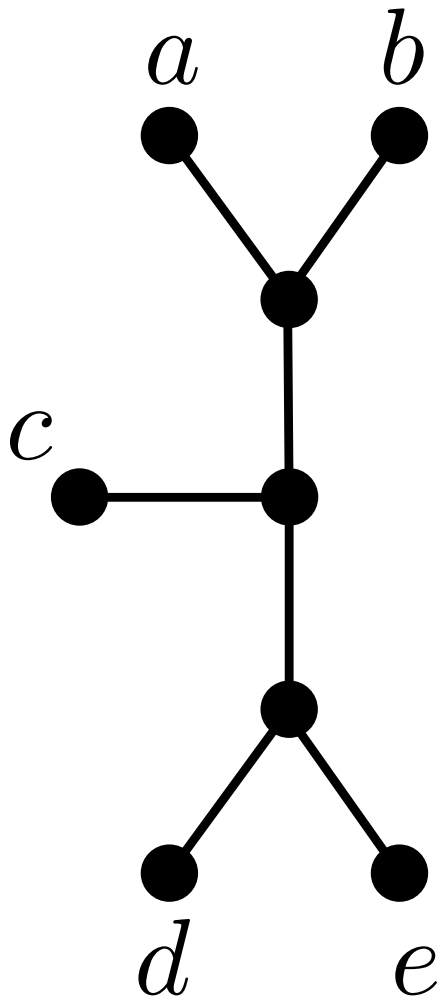


Phylogeny

A *phylogeny* $\mathcal{T}$ is a tree with

1. leaves labeled by a set X of *taxa* (or *species*), and
2. no degree two nodes.

Here, $\mathcal{T}$ is *unrooted*, without *branch lengths*.



Qualitative (or Discrete) Character χ

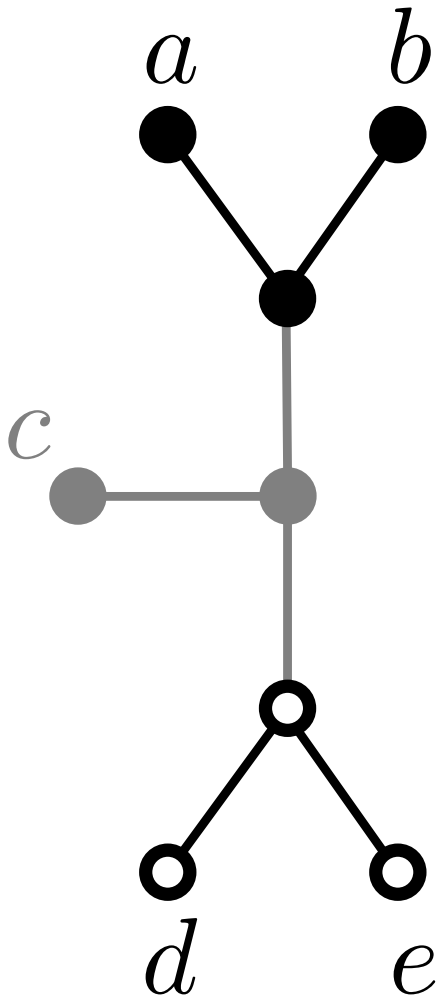
Grouping (partition) of taxa

- Each taxon takes on a discrete *state* of χ (or is *missing data*).
- Formally, a fn $\chi : X' \to \{1, 2, \ldots, r\}$ for some $X' \subseteq X$.
- Often just write $\chi = A_1|A_2|\ldots|A_r$, where $A_i = \chi^{-1}(i)$.
- A_i is a *state* of χ.

- In supertree context, derived from edges of phylogenies (edge splits taxa set into two).

Compatibility with a Phylogeny

	χ_1 ✓	χ_2
a	1	1
b	1	0
c	?	1
d	0	0
e	0	1

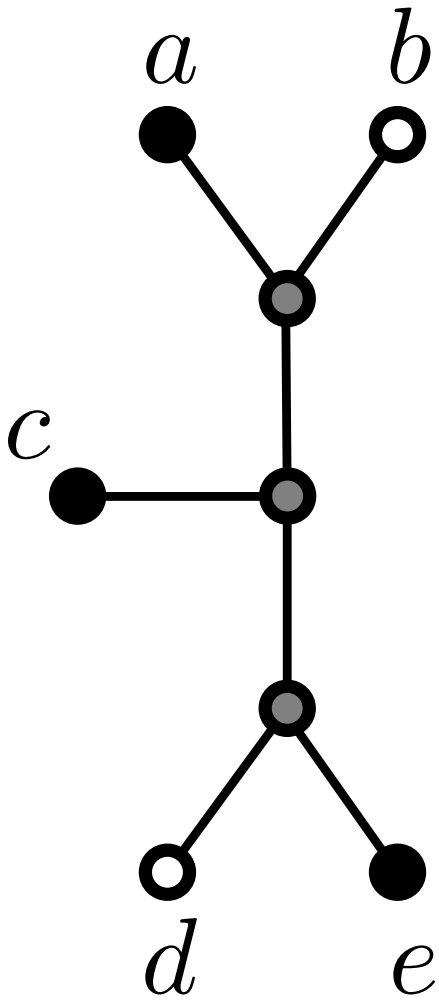


χ_1 Compatible

State-trees, black for state 1 and white for state 0, have no overlap.

Compatibility with a Phylogeny

	χ_1 ✓	χ_2 ✗
a	1	1
b	1	0
c	?	1
d	0	0
e	0	1



χ_2 Incompatible

State-trees overlap, sharing the three gray nodes.

Perfect Phylogeny Problems

The Character Compatibility (or Perfect Phylogeny) Problem (CC)

Is a given set of qualitative characters $\mathcal{C}$ that are missing data compatible with at least one phylogeny (i.e. has a perfect phylogeny)?

- Efficient reduction to a triangulation problem (Buneman '74, Steel '92)
- NP-hard when data is missing, i.e. supertree characters (Steel '92)
- Polynomial-time algorithms for special cases (Agarwala & Fernández-Baca '93; McMorris, Warnow & Wimer '93)
- Triangulation based ILP (Gusfield '10)

Perfect Phylogeny Problems

The Maximum Character Compatibility Problem (MC)

Given a set of qualitative characters $\mathcal{C}$ that are missing data, what is the largest (weighted or unweighted) subset of characters that has a phylogeny compatible with these characters?

- For $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ we are seeking the most (or highest weight set) of compatible edges.
- Efficient reduction to a triangulation problem (Bordewich, Huber & Semple '05; Gysel & Gusfield '11)
- NP-hard even with complete data (Day & Sankoff '86).
- ILP (Gusfield, Frid & Brown '07; Stevens & Gusfield '10)
- Triangulation based ILP (Gysel & Gusfield '11) and heuristics (Gysel, Stevens & Gusfield '13)

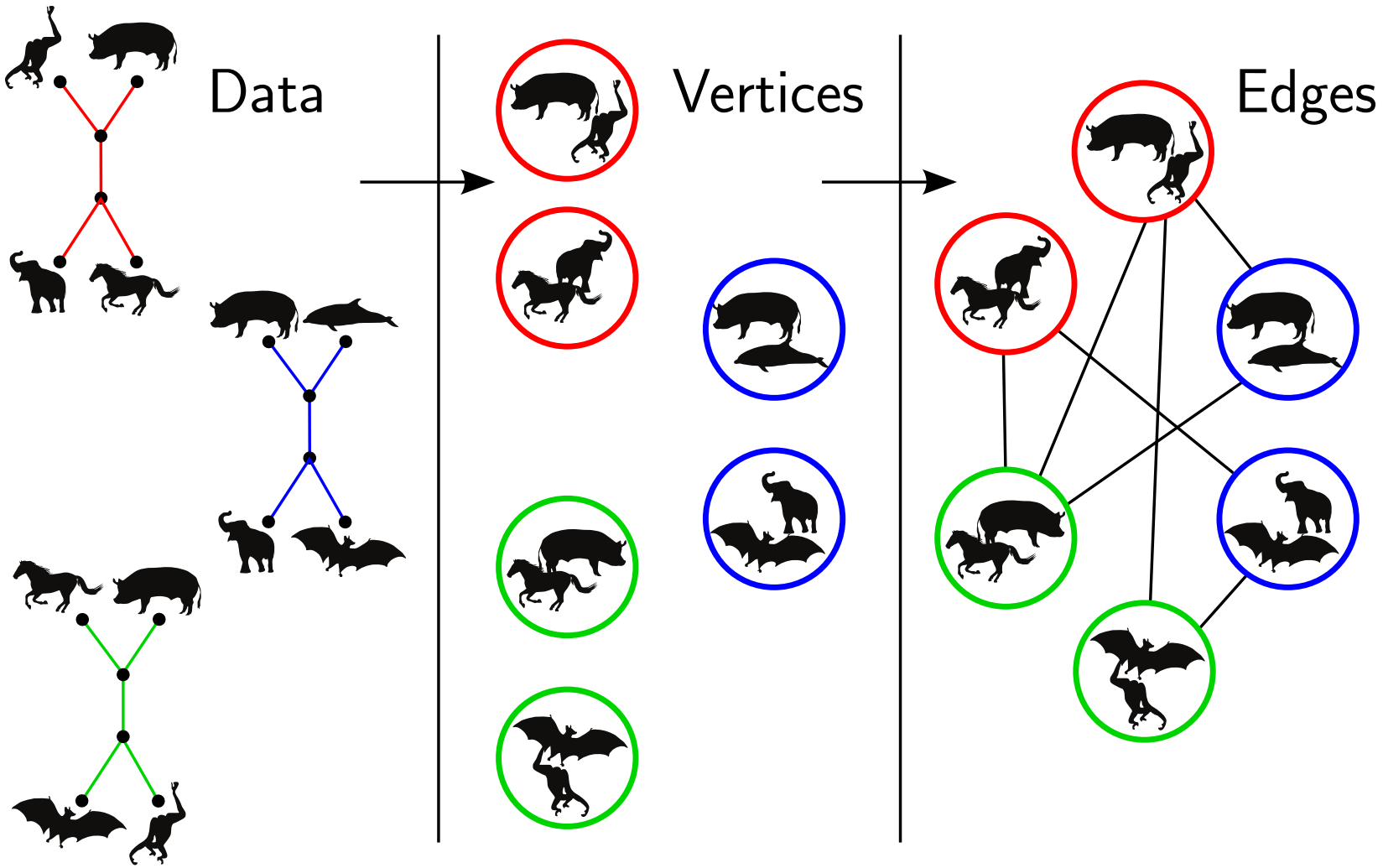
Perfect Phylogeny Problems

The Unique Perfect Phylogeny Problem (UP)

Given a set of qualitative characters $\mathcal{C}$ that are missing data, is there a unique phylogeny compatible with $\mathcal{C}$?

- Efficient reduction to a triangulation problem (Semple & Steel '02).
- If compatible phylogeny $\mathcal{T}$ given for X, determining if it is unique is CoNP-complete (Habib & Stacho '13; Bonet, Linz & St. John '12)

The Partition Intersection Graph $\mathrm{int}(\mathcal{C})$



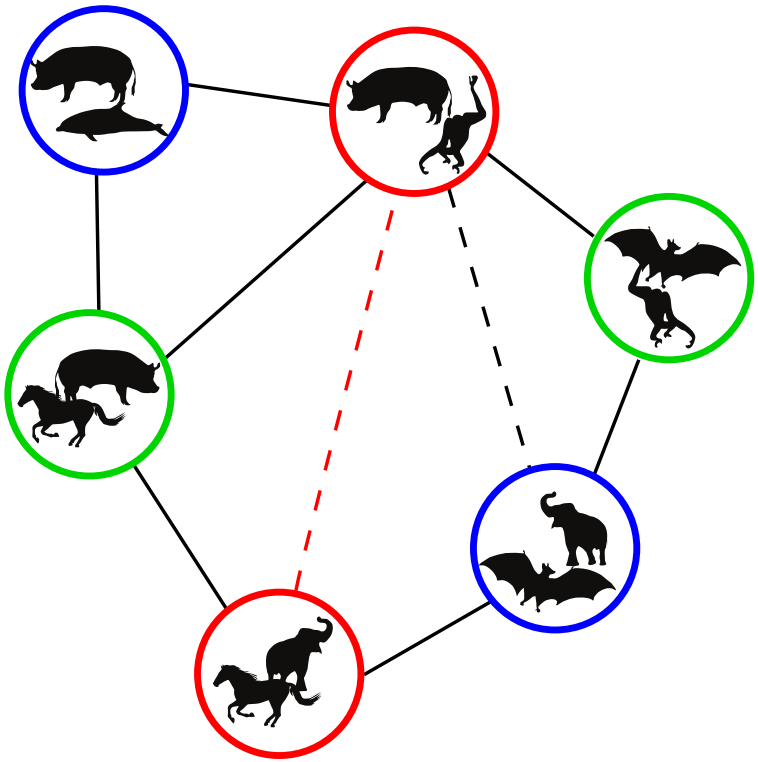
The Partition Intersection Graph $\mathrm{int}(\mathcal{C})$

Definition

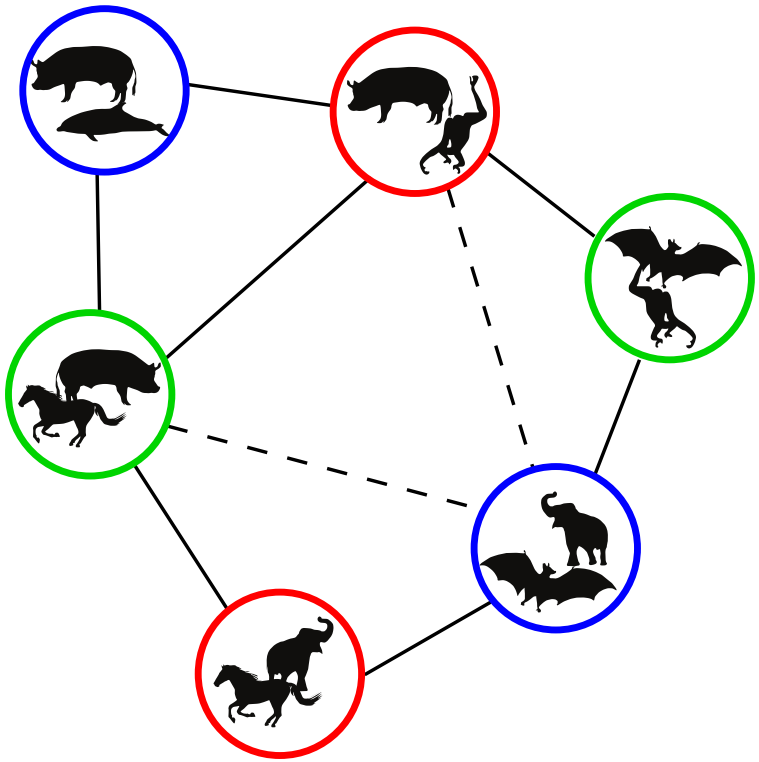
The *partition intersection graph* $\mathrm{int}(\mathcal{C})$ of a set of phylogenetic characters $\mathcal{C}$ has:

1. Vertex set $\{(A_i, \chi) \mid \chi \in \mathcal{C}$ and A_i is a state of $\chi\}$, and
2. Edges defined by intersection (i.e. (A_i, χ) and (A_j, χ') are an edge iff $A_i \cap A_j \neq \emptyset$)

Triangulations of $\mathrm{int}(\mathcal{C})$



Red character broken



Compatible (preserves coloring)

Triangulations of $\mathrm{int}(\mathcal{C})$

Triangulation Definitions

1. For an undirected graph $G = (V, E)$, a graph $H = (V, E \cup F)$ is a *triangulation* of G if it has no *chordless cycles* (*holes*) of length four or more.
2. H is *minimal* if no $F' \subset F$ yields a triangulation $H' = (V, E \cup F')$.
3. F are called *fill edges*.

Breaking Characters

A fill edge whose endnodes are derived from the same character, e.g. $f = (A, \chi)(A', \chi)$, *breaks* χ. We call f a *mono-chromatic fill edge*.

CC and MC as Triangulation Problems

Theorem: Character compatibility (CC) (Buneman '74, Steel '92)

Equivalent to finding a *proper* (minimal) triangulation of $\mathrm{int}(\mathcal{C})$ (i.e. has no mono-chromatic fill edges).

Theorem: Maximum character compatibility (MC)

1. (Bordewich, Huber & Semple '05) Equivalent to finding a triangulation of $\mathrm{int}(\mathcal{C})$ that breaks the fewest number of characters.
2. (Gysel & Gusfield '11) Suffices to consider minimal triangulations. Extension to weighted characters.

CC and MC as Triangulation Problems

Theorem: Character compatibility (CC) (Buneman '74, Steel '92)

Equivalent to finding a *proper* (minimal) triangulation of $\mathrm{int}(\mathcal{C})$ (i.e. has no mono-chromatic fill edges).

Theorem: Maximum character compatibility (MC)

1. (Bordewich, Huber & Semple '05) Equivalent to finding a triangulation of $\mathrm{int}(\mathcal{C})$ that breaks the fewest number of characters.
2. (Gysel & Gusfield '11) Suffices to consider minimal triangulations. Extension to weighted characters.

Theorem: Two-state MC (e.g. Supertree characters)

Equivalent to finding a minimal triangulation of $\mathrm{int}(\mathcal{C})$ with the fewest possible (or smallest weight) mono-chromatic fill edges.

A Characterization of Minimal Triangulations

Definition

A set $S \subseteq V$ is a *minimal separator* of G if S separates some vertex pair, but no proper subset separates it.

Theorem (Parra & Scheffler '97)

Let $G = (V, E)$ be an undirected graph.

1. Every minimal triangulation of G is obtained by *saturating* (adding edges to make a clique) a maximal set of *parallel* minimal separators.
2. Conversely, picking a set of minimal separators in this way and saturating them yields a minimal triangulation.

Remark (Gusfield '10)

Data generated by the coalescent simulator *ms* often results in $\mathrm{int}(\mathcal{C})$ with few minimal separators.

UP as a Triangulation Problem

Theorem (Semple & Steel '02)

$\mathcal{C}$ has a unique perfect phylogeny if and only if

1. $\mathrm{int}(\mathcal{C})$ has a unique proper minimal triangulation H, and
2. phylogeny derived from H is binary and its edges "distinguished" by $\mathcal{C}$.

Theorem

Let S^* be the set of minimal separators that appear in a proper minimal triangulation of $\mathrm{int}(\mathcal{C})$.

1. $\mathrm{int}(\mathcal{C})$ has a unique proper minimal triangulation if and only if S^* is a maximal set of parallel minimal separators of $\mathrm{int}(\mathcal{C})$.
2. UP algorithm: Compute S^*, check if parallel, compute compatible phylogeny, check its structure.

Note: Intractability of UP lies in computing S^*.

Weighted minimum fill

Weighting triangulations of $G = \mathrm{int}(\mathcal{C})$

Suppose characters weighted by w.

1. Weight monochromatic fill edges by w (or 1), other fill edges have weight 0
2. Weight a triangulation H of G by $w(H) = \sum_f w(f)$

- There is a proper minimal triangulation $\iff$ there is a minimal triangulation H with $w(H) = 0$
- A minimal triangulation H is optimal for MC $\iff$ it has minimum weight $w(H)$

Potential Maximal Cliques

Definition

$K \subseteq V$ is a *potential maximal clique* there is a minimal triangulation of G that has K as a maximal clique.

Treewidth and Minimum-Fill

Dynamic programming using potential maximal cliques and minimal separators solves *treewidth* and *minimum-fill* in:

- $O(|V|^3(\#\text{minseps})^3 + |V|^2|E|(\#\text{minseps})^2)$ time (Bouchitté & Todinca '02);
- improved to $O(|V|^2|E|(\#\text{minseps})^2)$ (Fomin et. al '08).

Triangulating Blocks

Definition

A *full block* of G is a pair (S, C) such that S is a minimal separator of G and C is a connect component of $G - S$ s.t. $N(C) \neq S$

Theorem (Bouchitté and Todinca '01)

K is a potential maximal clique if and only if the connected components C_i of $G - K$ satisfy:

1. $S_i = N(C_i)$ is a proper subset of K (i.e. is not full), and
2. saturating every S_i turns K into a clique.

Further, each S_i is a minimal separator of G ((S_i, C_i) is a full block of G).

Triangulating Blocks

Definition

The *realization* $R(S, C)$ of (S, C) is obtained by adding fill edges to $G[S \cup C]$ in order to saturate S (i.e. make S a clique)

Theorem

H is a minimal triangulation of $R(S, C)$ if and only if there is a potential maximal clique K such that

1. $S \subset K \subseteq S \cup C$,
2. $S \cup C$ is partitioned by K and connected components $C_1, \ldots, C_k$ of $R(S, C) - K$, and
3. $H[N(C_i) \cup C_i]$ is a minimal triangulation of $R(N(C_i), C_i)$.

Recursion Overview

Lemma (Kloks, Kratsch, Spinrad '97)

$$\text{mfi}(G) = \min_{S}(\text{fill}(S) + \sum_{C} \text{mfi}(R(S, C)))$$

Lemma (Bouchitté and Todinca '01)

$$\text{mfi}(R(S, C)) = \min_{S \subset K \subseteq S \cup C}[\text{fill}(K) - \text{fill}(S) + \sum \text{mfi}(R(S_i, C_i))]$$

Computing Potential Maxcliques and Minimal Separators

Minimal Separators of G

1. $O(|V|^3 \#\text{minseps})$ time (Berry et. al '00)
2. $O(1.6181^{|V|})$ time (Fomin & Villanger '12)

Potential Maximal Cliques of G

1. $O(|V|^2|E|(\#\text{minseps})^2)$ time (Bouchitté & Todinca '02)
2. $O(1.7549^{|V|})$ time (Fomin & Villanger '12)

Recursion Overview

Finding min-weight triangulation $\mathrm{mfi}_w(G)$

1. Compute full blocks (S, C)
2. From smallest to largest block (S, C):
   - Base case: saturate (S, C) if minimal, set $\mathrm{mfi}_w(S, C)$ to saturation cost $\mathrm{fill}_w(S \cup C)$
   - Else $\mathrm{mfi}_w(S, C) = \min_K [w(K) - w(S) + \sum_i w(R(S_i, C_i))]$
3. Another pass over minimal separators to compute $\mathrm{mfi}_w(G)$ and $\mathcal{S}^*$

Complexity Results

Character Compatibliity and Unique Perfect Phylogeny

Given m characters, r states, n taxa, these problems are solvable in $O(nm^2 + (rm)^4(\#\text{minseps})^2)$ time

Two-State Maximum Character Compatiblity

Given m characters, n taxa, MC is solvable in $O(nm^2 + m^4(\#\text{minseps})^2)$ time

Summary

1. Perfect phylogeny / character compatibility motivated by supertree estimation
2. Reduces to triangulating partition intersection graph
3. Algorithms are polynomial in number of minimal separators

Open Questions

1. Is it possible to count or approximate the number of minimal separators?
2. Is there a better potential maximal clique generating algorithm?

Thanks for listening!