CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO
College of Natural Sciences

Identifiability in Phylogenetic Networks under the Coalescent

New Directions in Algebraic Statistics

Hector Baños
Department of Mathematics
Wednesday, July 23, 2025

Joint Work

C. Áne

J. Xu

J. Rhodes

E. Allman

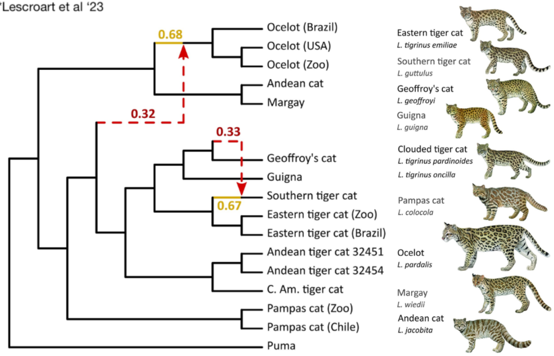
J. Mitchell

M. Garrote-Lopez

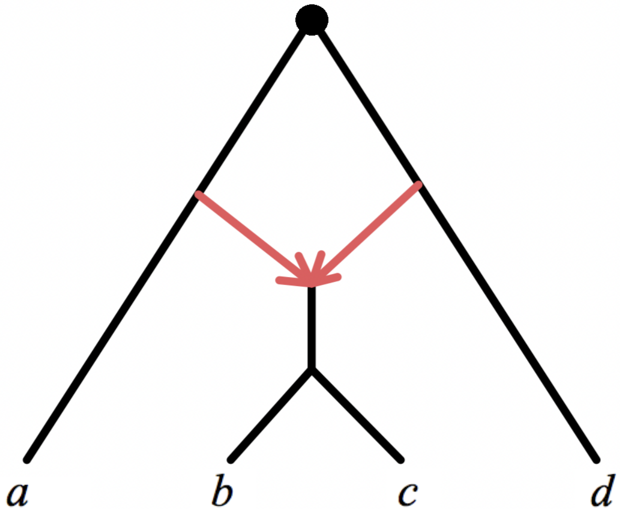
Species Networks (Admixture graphs)

- Phylogenetics is the study of the evolutionary history and relationships of organisms.
- New evidence shows hybridization has significantly influenced evolution
- Phylogenetic networks show evolutionary histories in the presence of hybridization.

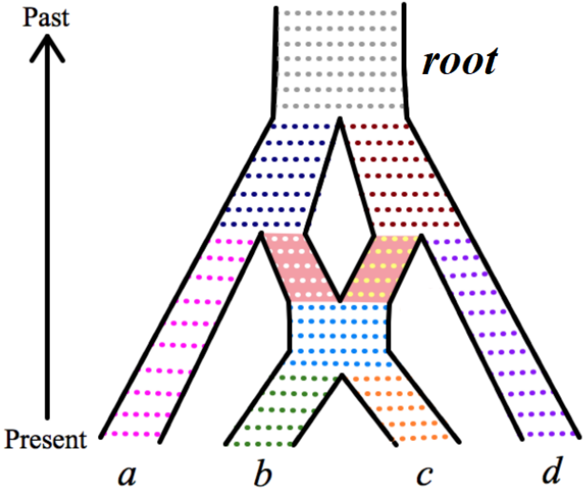
*Lescroart et al '23



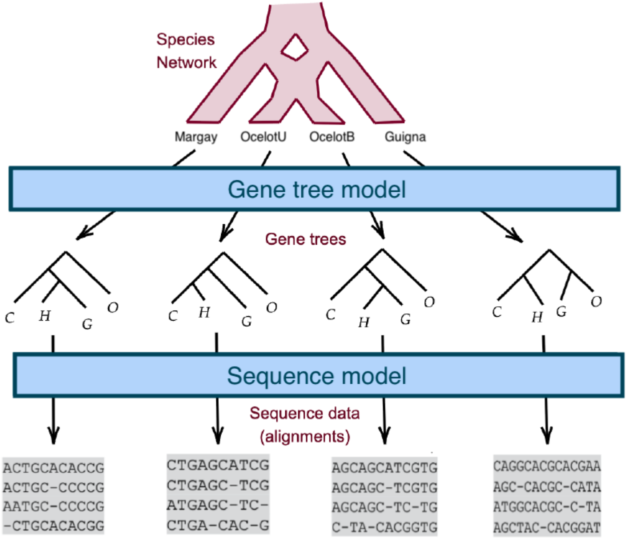
Species Network



Species Network



The model



Species
Network
Margay
OcelotU
OcelotB
Guigna
Gene tree model
Gene trees
C
H
G
O
C
H
G
O
C
H
G
O
C
H
G
O
Sequence model
Sequence data
(alignments)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

Data and Gene Tree Models

Data types

- Quartet concordance factors (CFs).
- Log-Det distances.
- Average genetic distances.
- Frequencies of full gene trees, or full site patterns.
- f_4 statistics.

Gene Tree Models

- **Network Multispecies Coalescent**: common or independent inheritance at hybrids.
- Displayed Tree: gene trees displayed in the network (no coalescent).

Literature

Network Multispecies Coalescent:

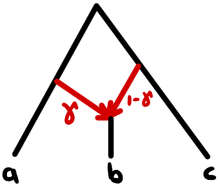
- Solís-Lemus & Ané 2016
- B. 2019
- Allman, B., & Rhodes 2022
- Allman, B., Mitchell, & Rhodes 2023
- **Allman,** B.**, Garrote-Lopez, & Rhodes** 2024
- **Rhodes,** B.**, Xu, & Ané** 2025
- **Allman, Ané,** B.**, & Rhodes** 2025

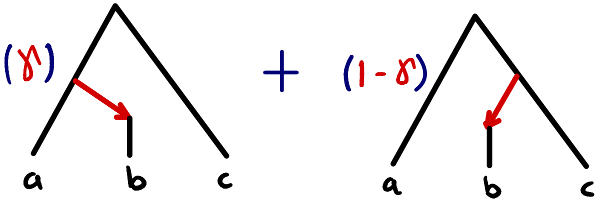
Displayed tree:

- Gross et al. 2021
- Hollering & Sullivant 2021
- Xu & Ané 2023
- Englander, Frohn, Gross, Holtgrefe, Van Iersel, Jones, & Sullivant 2025

The Displayed Tree Model

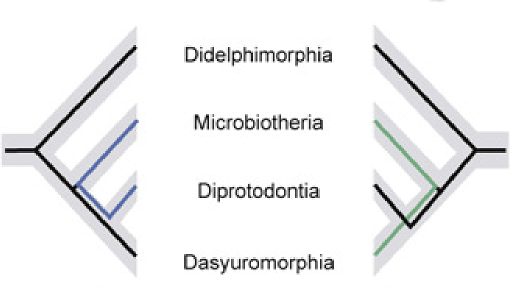
The Displayed Tree model assumes sequence evolve along the trees displayed by a network



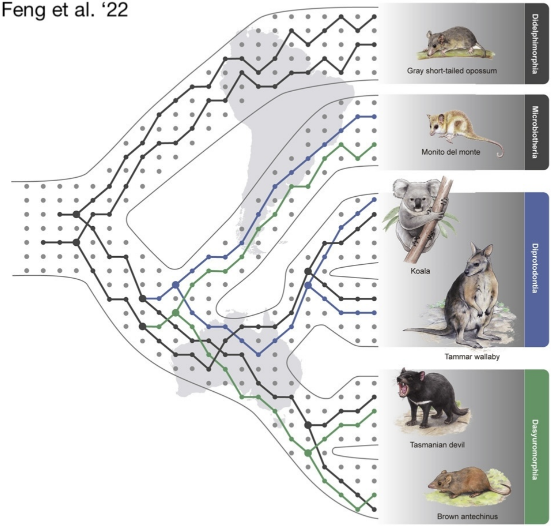


Why the Coalescent Model?

Incomplete lineage sorting (ILS)

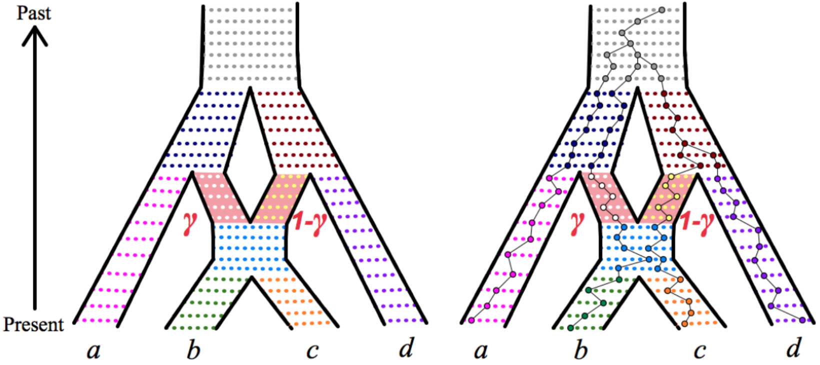


Feng et al. '22



The Network Multispecies Coalescent Model

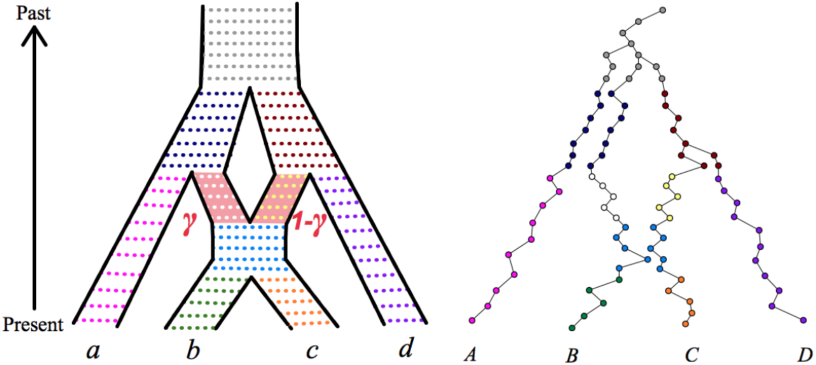
Kubatko & Meng ’09 - Degnan, Yu, & Nakhleh ’12 - Fogg, Ané, & Allman '24



The network multi-species coalescent describes a stochastic model of gene tree generation.

The Network Multispecies Coalescent Model

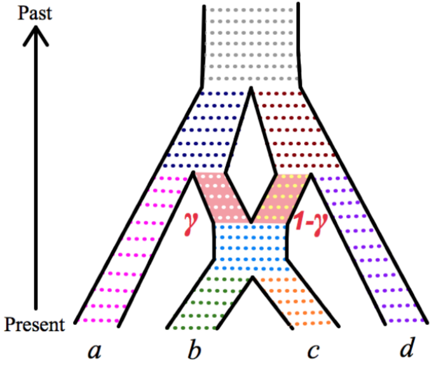
Kubatko & Meng ’09 - Degnan, Yu, & Nakhleh ’12

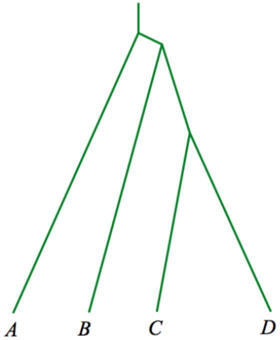


The network multi-species coalescent describes a stochastic model of gene tree generation.

The Network Multispecies Coalescent Model

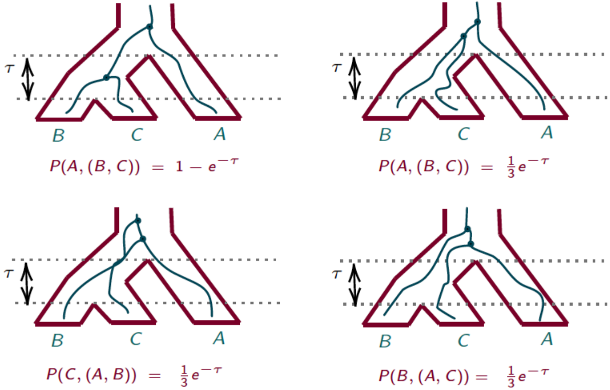
Kubatko & Meng ’09 - Degnan, Yu, & Nakhleh ’12





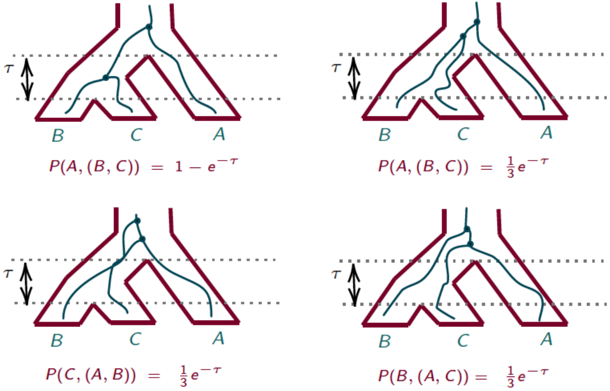
The network multi-species coalescent describes a stochastic model of gene tree generation.

The Network Multispecies Coalescent Model



$$P(A,(B,C)) = 1 - \frac{2}{3}e^{-\tau} \qquad P(C,(A,B)) = \frac{1}{3}e^{-\tau} \qquad P(B,(A,C)) = \frac{1}{3}e^{-\tau}$$

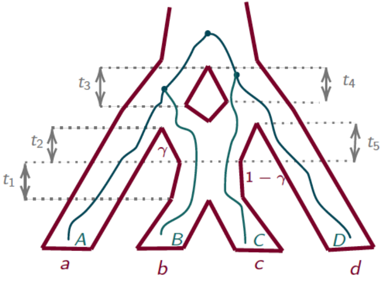
The Network Multispecies Coalescent Model



$$P(A,(B,C)) = 1 - \frac{2}{3}\mathbf{t} \qquad P(C,(A,B)) = \frac{1}{3}\mathbf{t} \qquad P(B,(A,C)) = \frac{1}{3}\mathbf{t}$$

t: *edge probability*

The Network Multispecies Coalescent Model



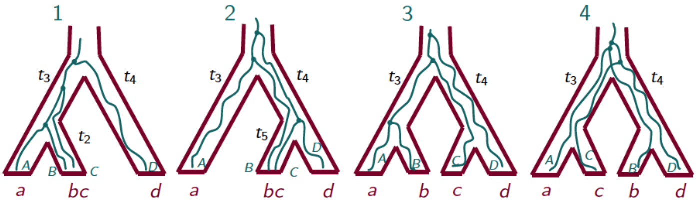
$$P((A,B),(C,D)) =$$

$$t_1\gamma^2 P_1 + t_1(1-\gamma)^2 P_2 +$$

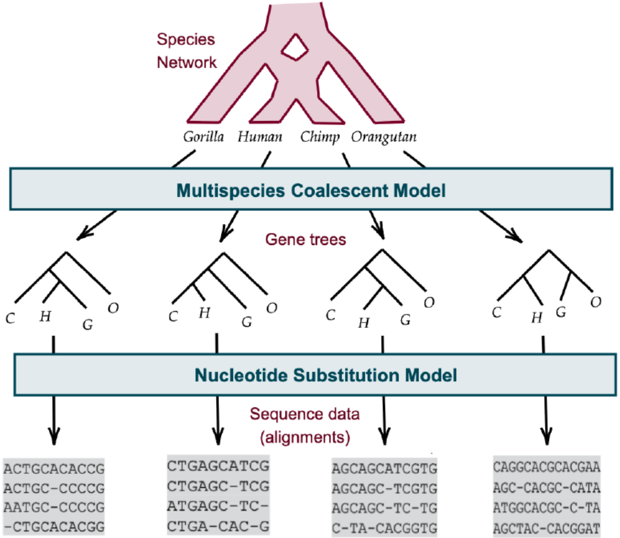
$$t_1\gamma(1-\gamma)P_3 + t_1\gamma(1-\gamma)P_4$$

$$P_1 = t_2 \cdot \frac{1}{3}$$

P_i is the probability of observing $((A,B),(C,D))$ under the MSC on tree **i**.



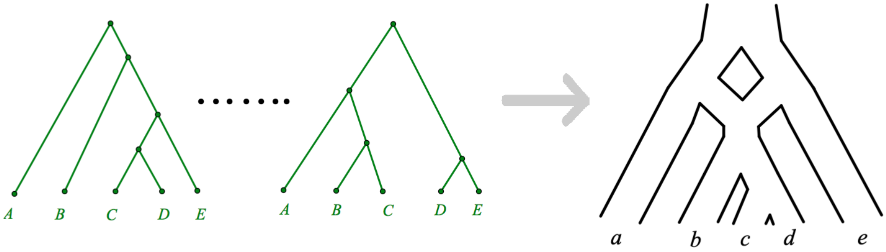
The model



Species
Network
Gorilla
Human
Chimp
Orangutan
Multispecies Coalescent Model
Gene trees
C
H
G
O
C
H
G
O
C
H
G
O
C
H
G
O
Nucleotide Substitution Model
Sequence data
(alignments)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

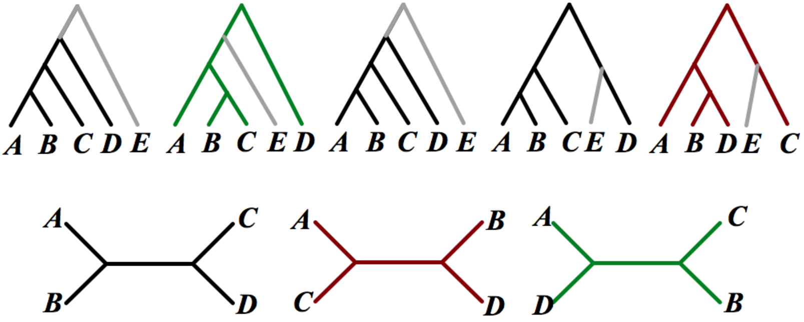
Identifying a Species Network from Gene Trees

Motivation: Given estimated gene trees sampled from the Network Multispecies Coalescent model (NMSC) on a network, identify properties of the network.



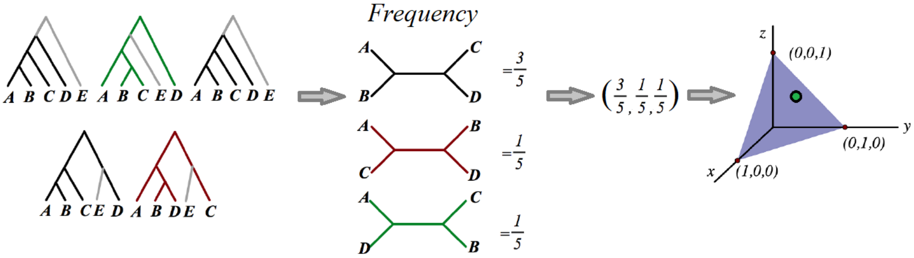
Quartet Frequencies

Given a sample of gene trees, one can calculate the *quartet frequencies* for any subset of four taxa.



$$Freq(AB|CD)=\frac{3}{5} \qquad Freq(AC|BD)=\frac{1}{5} \qquad Freq(AD|BC)=\frac{1}{5}$$

Quartet Concordance Factors



⋆ The **quartet Concordance Factor** for a set of 4 taxa a, b, c, d (denoted CF_{abcd}), is the vector of probabilities that a gene tree displays each possible quartet on the taxa.

- CFs are **polynomials** in terms of the parameters t_i and γ_i on a network.
- Quartet Frequencies are estimates of the CFs

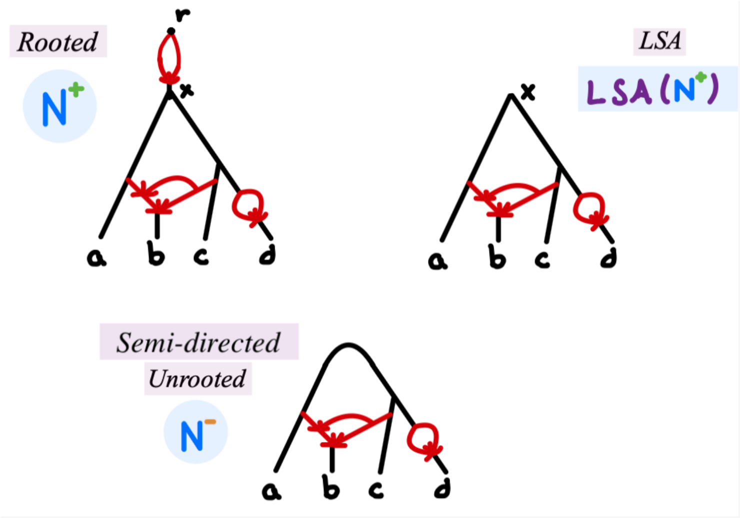
Quartet Concordance Factors

The quartet *CF*s for a topological semidirected network *N* define a **polynomial map**:

$$
\begin{array}{rccc}
CF(N): & \Theta(N) & \to & \overbrace{\triangle_2 \times \cdots \times \triangle_2}^{\binom{n}{4}} \subset \mathbb{C}^{3\binom{n}{4}} \\
 & (t_i, \gamma_j) & \mapsto & \left(\overline{CF}_{1234}, \ldots, \overline{CF}_{n-3,n-2,n-2,n}\right)
\end{array}
$$

- We denote by $\mathcal{V}(N) = \overline{\text{Im CF}}$ the variety of CF's associated to N.
- The set of multivariate polynomials in the CFs that vanish on the image of the parameterization forms an ideal, denoted $\mathcal{I}(N)$.
- Elements of $\mathcal{I}(N)$ are called invariants.

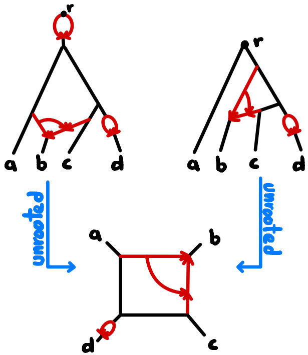
Semidirected Networks



The root is not identifiable from CFs

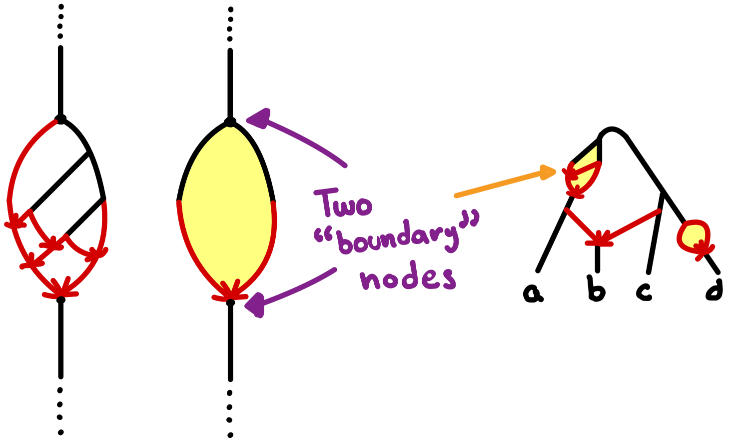
Theorem (Rhodes, B., Xu, & Áne)

Let N_1^+ and N_2^+ be two metric rooted networks on a set X. If $N_1^- = N_2^-$, then for every 4-taxon set $\mathrm{CF}(N_1^+) = \mathrm{CF}(N_2^+)$. In particular, the subgraph above the LSA of a rooted network does not affect quartet CFs.



The 2 sub-blobs of a network

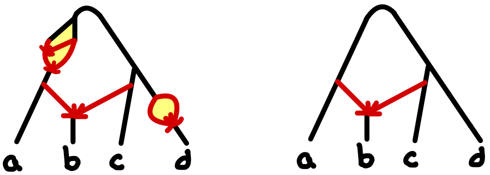
A 2 sub-blob:



2-blobs are not identifiable from CFs

Theorem (Rhodes, B., Xu, & Áne)

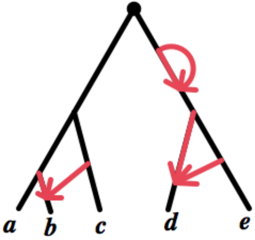
Let N be a metric network and G be a 2-sub-blob in N with boundary nodes u and v. Then there exists $t = t(G, \rho) \geq -\log(3/2)$ such that replacing G with a single tree edge (u, v) or (v, u) of length t leaves the quartet concordance factors of N unchanged. If G does not trap the root, or if u (or v) has a single descendant leaf in $N \setminus \{v\}$, then $t \geq 0$.



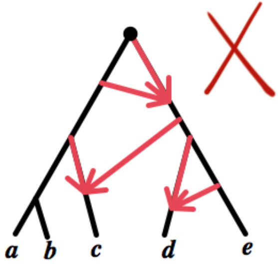
So what is identifiable?????

Definition

A network $\mathcal{N}$ is **level-1** if no pair of cycles in $\mathcal{N}$ share an edge.



level-1



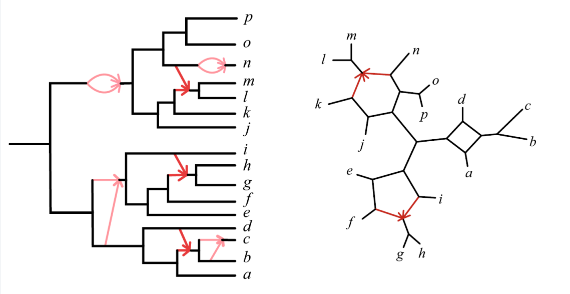
Not level-1

Level-1 Network Identifiability

Following Solís-Lemus & Ané '16.

Theorem (B.)

Let N be a rooted binary metric level-1 species network. Let N' be the semidirected topological network obtained from N by contracting all 2- and 3-cycles, and undirecting the hybrid edges in 4-cycles. Then, under the NMSC model, from N's quartet CFs the network N is identifiable.



Level-1 Network Identifiability

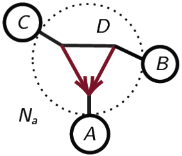
CSUSB

Proposition (Allman, B., Garrote-Lopez, & Rhodes)

Let N be a semidirected networks (not necessarily level-1) with an undirected structure as in the figure or the network with the 3-cycle shrunk to a node. Then

$$D = \begin{cases} \text{a node} & \text{if } G_{abc} = G_{bca} = G_{cab} = 0, \\ \text{a 3-cycle with } A & \text{if } G_{abc} > 0,\ G_{bca} \leq 0,\ G_{cab} \leq 0 \\ \quad \text{below the hybrid node} & \\ \text{a 3-cycle with } A \text{ or } B & \text{if } G_{abc} > 0,\ G_{bca} > 0,\ G_{cab} < 0 \\ \quad \text{below the hybrid node} & \end{cases}$$

where $G_{xyz} = CF_{xz|xz} CF_{xy|yz} - 2CF_{yz|yz} CF_{xy|xz} + CF_{xy|xy} CF_{xz|yz}$.



Level-1 Network Identifiability

Theorem (Allman, B., Garrote-Lopez, & Rhodes)

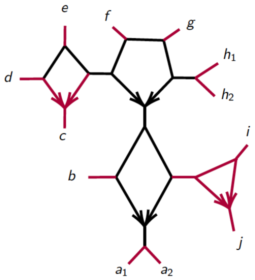
Let N be a level-1 metric binary semidirected network with no 2-cycles. Then from quartet CFs all numerical parameters on N are identifiable except:

Level-1 Network Identifiability

Theorem (Allman, B., Garrote-Lopez, & Rhodes)

Let N be a level-1 metric binary semidirected network with no 2-cycles. Then from quartet CFs all numerical parameters on N are identifiable except:

1. *pendant edge lengths,*
2. *hybrid edge lengths when the hybrid node has exactly one descendant taxon,*
3. *for 3-cycles, hybridization parameters and the lengths of the six edges in and incident to the cycle,*
4. *for 4-cycles, the hybridization parameter and edge lengths of edges adjacent to the hybrid node as in the previous slide.*



Network Inference Algorithm - NANUQ

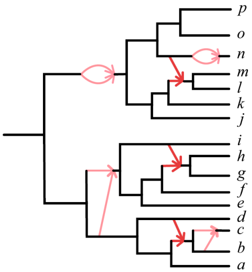
The NANUQ algorithm for inference of topological species networks[1].

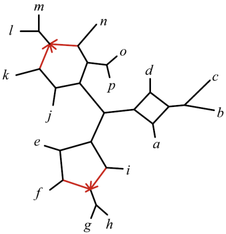
Input:

A collection of topological gene trees on a taxon set X, a hypothesis testing level α.

Ouput:

When the input comes from a level-1 rooted species network, the unrooted species network, after suppressing small cycles, and the directions of hybrid edges in 4-cycles.

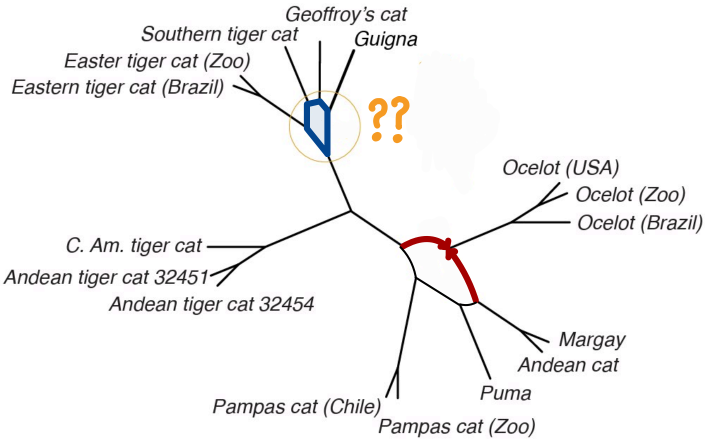




[1] NANUQ is the Inupiaq word for polar bear

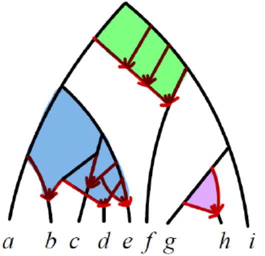
Level-1 might be too restrictive for empirical data

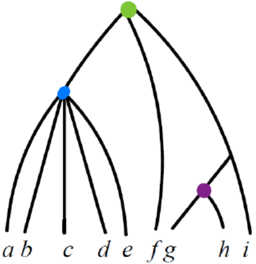
NANUQ on the Leopardus data:

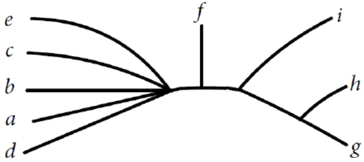


Beyond Level-1 - Tree of Blobs

The **tree of blobs** of a network is the tree obtained after contracting each “blob” to a node.





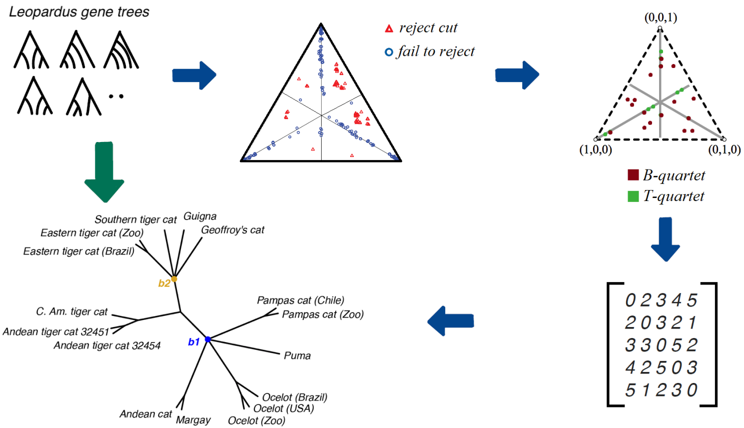


Theorem (Rhodes, B., Xu, & Áne)

*The tree of an arbitrary network is identifiable from CFs **(some additional requirements are needed)**.*

Tree of Blobs Inference Algorithm - TINNiK

Tree of blobs INference for a NetworK[2]



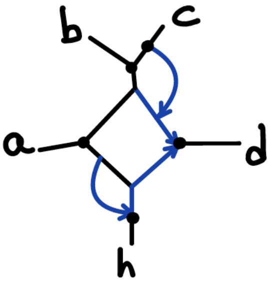
[2] Tinnik is the Inupiaq word for bearberry

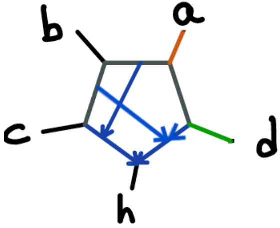
Outer-labeled Planar Blobs

Definition

A network $\mathcal{N}$ is outer-labelled planar (OLP) if it can be represented in the plane:

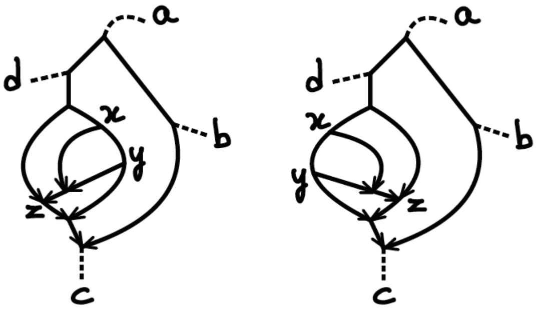
- with no edge crossing (planar), and
- with all taxa are in the "outside" (outer-labelled)





Circular Order

In an outer-labeled planar blob,the **circular order** of taxa is well defined.

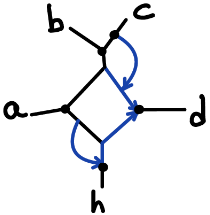


Different planar embedding must have *a*, *b*, *c*, *d* in the same order along the outer face.

Circular Order

Theorem (Rhodes, B., Xu, & Áne)

For a binary outer-labeled planar blob, the full circular order is identifiable from CFs.

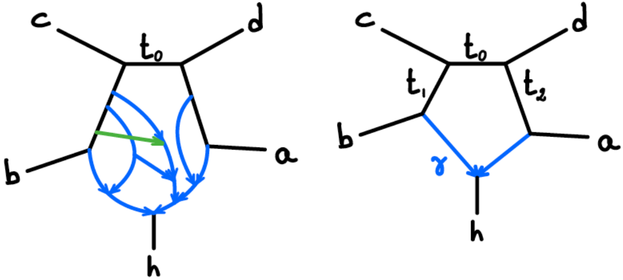


Along **Alexandr**, Coons, **Meshkat**, Long, & **Gross**, we are exploring different things of circular orders. Including developing an algorithm for its inference and looking at algebraic properties.

Indistinguishable

Not everything is honey over corn flakes....

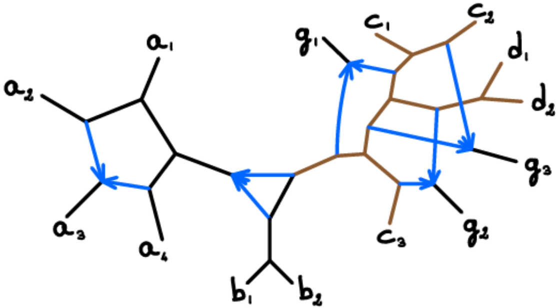
These are not distinguishable from CFs.

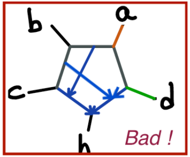


Where are we currently on identifiability?

Under CFs we have identifiability results for networks of **arbitrary level**, under the restriction that these are:

- Binary
- Galled
- Tree-child
- Class $\mathfrak{C}_4$ or
  (multiple samples
  per taxon needed)
- Class $\mathfrak{C}_5$





Thank you!

- *Beyond level-1: Identifiability of a class of galled tree-child networks.*
  ES Allman, C Ane, H Banos, JA Rhodes. Arxiv 2025.
- *Identifying circular orders for blobs in phylogenetic networks.*
  JA Rhodes, H Banos, J Xu, C Ané. Advances in Applied Mathematics 2025.
- *Identifiability of Level-1 Species Networks from Gene Tree Quartets.*
  ES Allman, H Baños, M Garrote-Lopez, JA Rhodes. BMAB 2024.

Hector Banos
hector.banos@csusb.edu
Department of Mathematics
California State University, San Bernardino

CALIFORNIA STATE UNIVERSITY SAN BERNARDINO
College of Natural Sciences

DMS-2331660