Phylogenetic Network Models and Graphical Models

Seth Sullivant

North Carolina State University

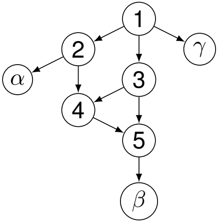
July 24, 2025

Motivation: Identifiability of Level-2 Networks

Theorem (Englander-Frohn-Gross-Holtgrefe-Van Iersel-Jones-S)

The network parameter of the displayed tree model under the Jukes-Cantor substitution is generically identifiable when the network parameter is an n-leaf binary, triangle-free, strongly tree-child, level-2 semi-directed phylogenetic network.

- Proof uses a range of tools.
    - Matroids, Phylogenetic Invariants/Ideals, Inequalities
- Challenges: Stacked reticulations, triangles



Graphical Models

- Graphical models are a flexible framework for building statistical models on (large) collections of random variables.
- Edges of different types represent different types of interactions between neighboring random variables.
  - directed edges: $i \to j$
  - bidirected edges: $i \leftrightarrow j$
  - undirected edges: $i - j$
- Graph is used to express both
  - conditional independence structures between random variables
  - parametric representations of the model.
- In this talk: directed acyclic graphs (DAGs) and discrete random variables.

Parametrization

- Let $G = (V, D)$ be a directed acyclic graph.
- For each $v \in V$, we have a discrete random variable X_v.
- For each $v \in V$, $\mathrm{pa}(v)$ is the parent set of v:

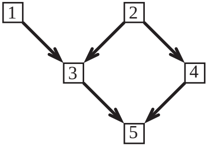
$$\mathrm{pa}(v) = \{u \in V : u \to v \in D\}.$$

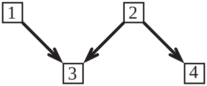
- DAG Graphical model expresses the joint distribution of $X = (X_v | v \in V)$ via a recursive factorization:

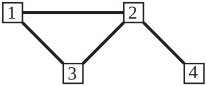
$$p(x) = \prod_{v \in V} p_v(x_v | x_{\mathrm{pa}(v)}).$$

Conditional Independence

- DAG models also specified by conditional independence structures
- $X_A \perp\!\!\!\perp X_B | X_C$ holds iff A and B are d-separated given C.





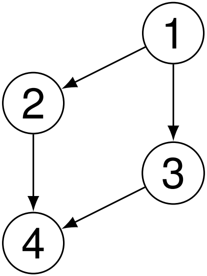


$\{1, 3\}$ and $\{4\}$ are d-separated given $\{2\}$.
So $(X_1, X_3) \perp\!\!\!\perp X_4 | X_2$ holds in this graph.

Theorem (Recursive factorization)

A probability distribution has a recursive factorization according to a DAG G if and only if it satisfies the global conditional independence statements of G.

Example: Directed 4-cycle



- Recursive factorization

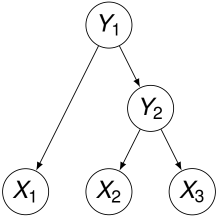
$$p(x_1, x_2, x_3, x_4) = p_1(x_1)p_2(x_2|x_1)p_3(x_3|x_1)p_4(x_4|x_2, x_3)$$

- Conditional independence

$$X_2 \perp\!\!\!\perp X_3 | X_1 \qquad X_1 \perp\!\!\!\perp X_4 | (X_2, X_3)$$

Phylogenetic Models

- Assuming site independence:
- Phylogenetic Model is a latent class graphical model
- Leaf $v \in T$ is random variable $X_v \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.
- Internal nodes $v \in T$ are latent random variables Y_v



$$p(x_1, x_2, x_3) = \sum_{y_1} \sum_{y_2} p_1(y_1) p_2(y_2|y_1) p_3(x_1|y_1) p_4(x_2|y_2) p_5(x_3|y_2)$$

Substitution Models

- Phylogenetic models are typically submodels of the hidden variable graphical model on a tree.
- This is obtained by specifying a structure on the substitution model/transition matrix structure.

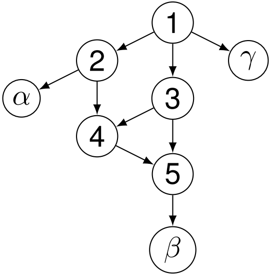
$$M^e = \begin{pmatrix} p_e(A|A) & p_e(A|C) & p_e(A|G) & p_e(A|T) \\ p_e(C|A) & p_e(C|C) & p_e(C|G) & p_e(C|T) \\ p_e(G|A) & p_e(G|C) & p_e(G|G) & p_e(G|T) \\ p_e(T|A) & p_e(T|C) & p_e(T|G) & p_e(T|T) \end{pmatrix}$$

- Equivariant models:
  - Let G be a subgroup of S_4, acting on $\{A, C, G, T\}$.
  - Equivariant: for all $g \in G$, $p_e(x|y) = p_e(g(x)|g(y))$

$$\begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix} \qquad \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix} \qquad \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}$$

Phylogenetic Networks

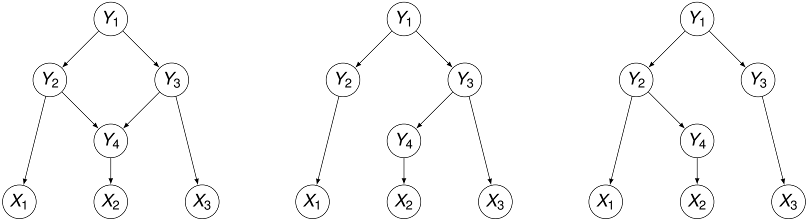
- Use a more complicated graph than a tree to represent evolutionary relationships between species.



- Tree vertex: One or fewer incoming edges
- Reticulation vertex: Two or more incoming edges
- Reticulations vertices are used to represent hybridization, gene transfer, or other non-tree-like evolution.

The Displayed Tree Model

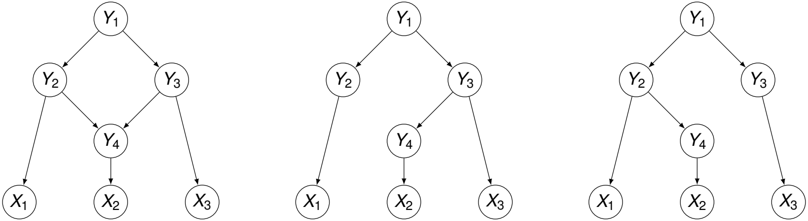
- Probability distribution for network obtained by weighted sum over all displayed trees of that network.
- Displayed trees: Choose one edge at each reticulation vertex



$$\begin{aligned} p(x_1,x_2,x_3) &= (1-\delta)\sum_y p_1(y_1)p_2(y_2|y_1)p_3(y_3|y_1)p_4(y_4|y_3)p_6(x_1|y_2)p_7(x_2|y_4)p_8(x_3|y_3) \\ &+ \quad \delta\sum_y p_1(y_1)p_2(y_2|y_1)p_3(y_3|y_1)p_5(y_4|y_2)p_6(x_1|y_2)p_7(x_2|y_4)p_8(x_3|y_3) \end{aligned}$$

- Note that transition matrices are reused in both trees. Same edge, same transition matrix.

The Displayed Tree Model as a DAG



$$\begin{aligned} p(x_1, x_2, x_3) &= (1-\delta)\sum_y p_1(y_1)p_2(y_2|y_1)p_3(y_3|y_1)p_4(y_4|y_3)p_6(x_1|y_2)p_7(x_2|y_4)p_8(x_3|y_3) \\ &+ \quad \delta\sum_y p_1(y_1)p_2(y_2|y_1)p_3(y_3|y_1)p_5(y_4|y_2)p_6(x_1|y_2)p_7(x_2|y_4)p_8(x_3|y_3) \\ &= \sum_y p_1(y_1)p_2(y_2|y_1)p_3(y_3|y_1)q(y_4|y_2, y_3)p_6(x_1|y_2)p_7(x_2|y_4)p_8(x_3|y_3) \end{aligned}$$

- $q(y_4|y_2, y_3)$ is a restricted version of the general conditional distribution $p(y_4|y_2, y_3)$.

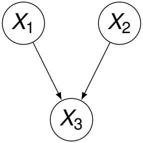
$$q(y_4|y_2, y_3) = (1-\delta)p_4(y_4|y_3) + \delta p_5(y_4|y_2)$$

Conditional Distributions from the DTM

Proposition

The displayed tree phylogenetic network model is the submodel of the DAG model where for each i:

$$p_i(x_i|x_{\mathrm{pa}(i)}) = \sum_{j \in \mathrm{pa}(i)} \delta_j p_{ji}(x_i|x_j).$$



- General Markov model
  - Full conditional distribution: $4^2(4-1) = 48$ parameters
  - Displayed tree conditional distribution: $2 \times 4(4-1) + 1 = 25$ parameters

Loss of Dimension in Conditional Distributions

- The dimension drops even more!

Proposition (Casanellas-Fernández Sánchez-Gross-Hollering-S)

The dimension of the General Markov model κ states, 2 parent reticulation conditional distribution has $1 + 2\kappa(\kappa - 1)$ parameters but only dimension

$$1 + 2\kappa(\kappa - 1) - \kappa$$

$$p_x(x|y,z) = \delta p_{yx}(x|y) + (1-\delta)p_{zx}(x|z)$$

$$c_{ijk} = \delta a_{ij} + (1-\delta)b_{ik}$$

$$\begin{array}{rcl} c_{ij_1k_1} + c_{ij_2k_2} & = & c_{ij_1k_2} + c_{ij_2k_1} \\ \delta a_{ij_1} + (1-\delta)b_{ik_1} + \delta a_{ij_2} + (1-\delta)b_{ik_2} & = & \delta a_{ij_1} + (1-\delta)b_{ik_2} + \delta a_{ij_2} + (1-\delta)b_{ik_1} \end{array}$$

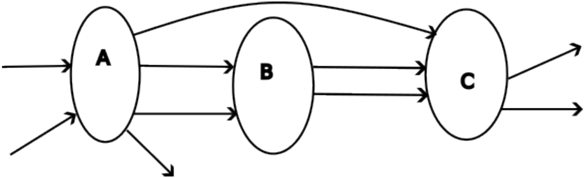
Local Structure in a DAG

Definition

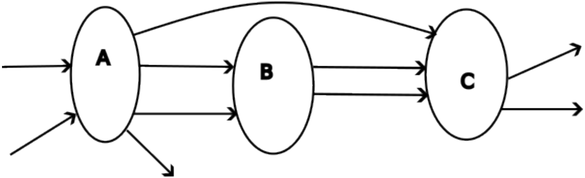
Let $G = (V, D)$ be a DAG. Let $A, B, C \subseteq V$ be disjoint with:

- For each vertex $b \in B$, every edge $i \to b$ has $i \in A \cup B$
- For each vertex $b \in B$, every edge $b \to i$ has $i \in B \cup C$
- For each vertex $c \in C$, every edge $i \to c$ has $i \in A \cup B \cup C$

We say that the triple of vertices (A, B, C) gives a local structure in G.



Local structures in DAGs



Proposition

Let $G = (V, D)$ be a DAG and (A, B, C) a local structure in G. Then

$$p(x_B, x_C | x_{\mathrm{an}(B \cup C)}) = p(x_B, x_C | x_A).$$

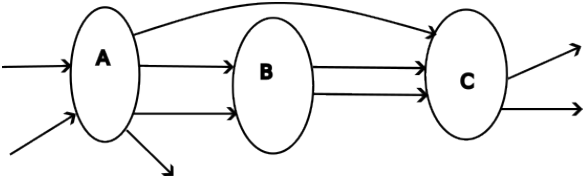
Local Modifications to DAGs

Definition

Let G be a DAG with a local structure (A, B, C). Let $V' = V \setminus (A \cup B \cup C)$. Let G' be a new DAG with vertex set $V' \cup A \cup B' \cup C$ that satisfies the following properties

- (A, B', C) is a local structure in G'.
- Let $i, j \in V' \cup A$. Then $i \to j \in G$ if and only if $i \to j \in G'$.
- Let $i \in C$ and $j \in V'$. Then $i \to j \in G$ if and only if $i \to j \in G'$.

The graphs G and G' are called local modifications of each other.

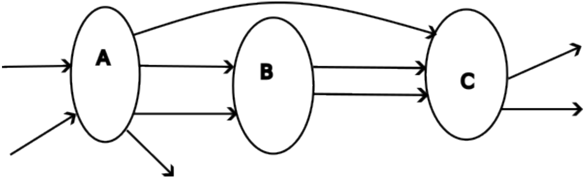


Local Modifications Theorem

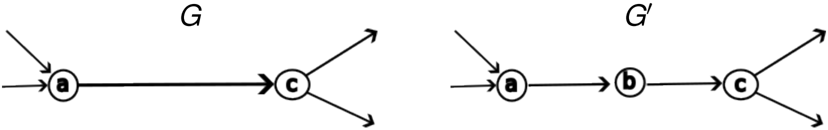
Theorem

- *Let G_1 and G_2 be two graphs that are local modifications of each other with local structures (A, B, C) and (A, B', C) respectively.*
- *Suppose that the family of conditional distributions in the two models $p_{G_1,C|A}(x_C|x_A)$ and $p_{G_2,C|A}(x_C|x_A)$ are the same.*
- *Suppose further that each of the other set of distributions $p_{i|\mathrm{pa}(i)}(x_i|x_{\mathrm{pa}(i)})$ is the same in both graphs.*

Then the family of joint distributions with the variables in X_B and $X_{B'}$ hidden variables are the same in both models.



Example: Subdividing an edge



- Subdividing an edge in a DAG is a local modification.

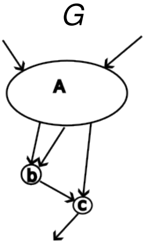
$$p_{G',c|a}(x_c|x_a) = \sum_{x_b} p_{G',c|b}(x_c|x_b)p_{G',b|a}(x_b|x_a)$$

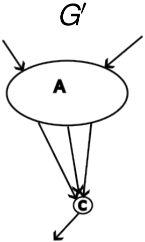
Proposition

The phylogenetic network model on G and G' give the same family of probability distributions if the set of model transition matrices is

- *closed under matrix multiplication*
- *splittable*.

Stacked Reticulations





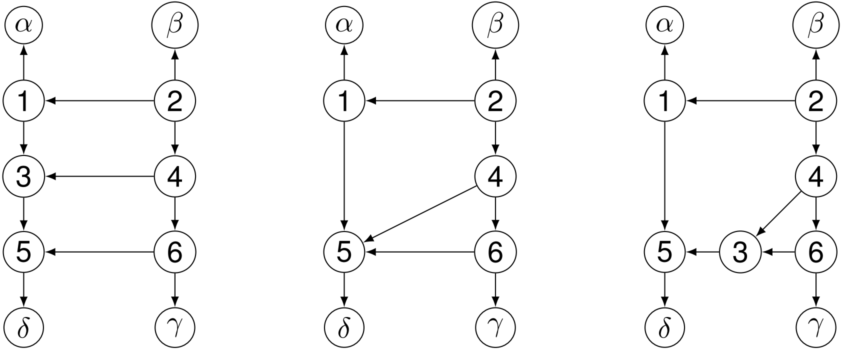
- Contracting a stacked reticulation is a local modification.

Proposition

The phylogenetic network model on G and G′ give the same family of probability distributions if the set of model transition matrices is

- *closed under matrix multiplication*
- *closed under convex combinations, and*
- *splittable.*

Stacked Reticulations are Unidentifiable



- All three networks give the same family of probability distributions on leaves $\alpha, \beta, \gamma, \delta$ under any equivariant phylogenetic model.
- Stacked reticulations are never identifiable under the displayed trees model.

Ranks of Flattenings

Theorem (Allman-Rhodes)

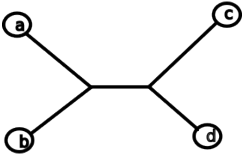
Let T be a tree, $\mathcal{M}^T$ a phylogenetic model on T with κ states. Let $A|B$ be a bipartition of the leaves of T.

- *If $A|B$ is a valid split of T, then for $P \in \mathcal{M}^T$*

$$\operatorname{rank} \operatorname{flat}_{A|B} P \leq \kappa$$

- *If $A|B$ is not a valid split of T, then for generic $P \in \mathcal{M}^T$*

$$\operatorname{rank} \operatorname{flat}_{A|B} P > \kappa$$



$$\operatorname{flat}_{ab|cd} P = \begin{pmatrix} p_{0000} & p_{0001} & p_{0010} & p_{0011} \\ p_{0100} & p_{0101} & p_{0110} & p_{0111} \\ p_{1000} & p_{1001} & p_{1010} & p_{1011} \\ p_{1100} & p_{1101} & p_{1110} & p_{1111} \end{pmatrix}$$

$$\operatorname{rank} \operatorname{flat}_{ab|cd} P \leq 2$$

Ranks of Flattenings

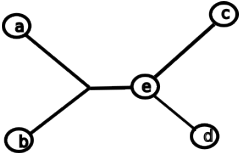
Theorem (Allman-Rhodes)

Let T be a tree, $\mathcal{M}^T$ a phylogenetic model on T with κ states. Let $A|B$ be a bipartition of the leaves of T.

- *If $A|B$ is a valid split of T, then for $P \in \mathcal{M}^T$*

$$\mathrm{rank}\,\mathrm{flat}_{A|B} P \leq \kappa$$

- This result follows from conditional independence in the tree, given hidden variables.



$$\mathrm{rank}\,\mathrm{flat}_{ab|cd} P \leq 2$$

$$(X_a, X_b) \perp\!\!\!\perp (X_c, X_d) | X_e$$

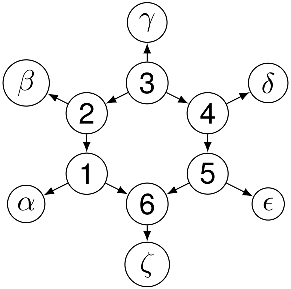
Ranks of Flattenings for Networks

Theorem (Casanellas-Fernández Sánchez-Gross-Hollering-S)

Let N be a network, and $\mathcal{M}^N$ the phylogenetic model on κ states. Let $A|B$ be a bipartition of the leaves.

- *$m_N(A|B)$ is the minimum number of edges separating A and B.*
- *$\ell_N(A|B)$ is the largest parsimony score of displayed trees in N.*

Then for generic $P \in \mathcal{M}^N$: $\quad \kappa^{\ell_N(A|B)} \leq \operatorname{rank} \operatorname{flat}_{A|B} P \leq \kappa^{m_N(A|B)}$.



$$m_N(\alpha\beta\gamma|\delta\epsilon\zeta) = 2$$

$$\ell_N(\alpha\beta\gamma|\delta\epsilon\zeta) = 2$$

- This result can be used to prove identifiability of level-1 networks via flattening ranks.

Equivariant DAGs?

- Equivariant tree models:
  - Let G be a subgroup of S_4, acting on $\{A, C, G, T\}$.
  - Equivariant: for all $g \in G$, $P(x|y) = P(g(x)|g(y))$
- Equivariant DAG models
  - For all $g \in G$, $P(x|y_1, \ldots, y_k) = P(g(x)|g(y_1), \ldots, g(y_k))$

$$\begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix} \qquad \left(\begin{array}{cccc|cccc|cccc|cccc} a & b & b & b & d & c & e & e & d & e & c & e & d & e & e & c \\ c & d & e & e & b & a & b & b & e & d & c & e & e & d & e & c \\ c & e & d & e & e & c & d & e & b & b & a & b & e & e & d & c \\ c & e & e & d & e & c & e & d & e & e & c & d & b & b & b & a \end{array}\right)$$

- The equivariant displayed tree model is a submodel of the equivariant DAG model.
- Maybe the equivariant DAG model is easier to study?

Summary and Conclusions

- Phylogenetic network models are used when non-tree-like structures are present in evolutionary histories.
- The displayed tree model is a submodel of the directed acyclic graphical model from the same network.
- We used this connection to show:
  - New nonidentifiability results for the displayed tree model with stacked reticulations
  - New ranks of flattening results for networks