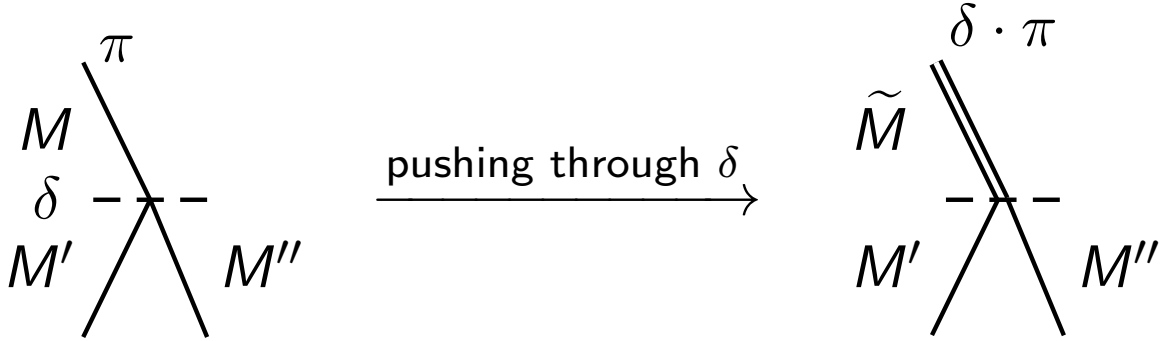
Distinguishing Convergence on Two-Taxon and Three-Taxon Networks

Jonathan Mitchell
Supervisors: Barbara Holland, Jeremy Sumner

University of Tasmania

November 6, 2014

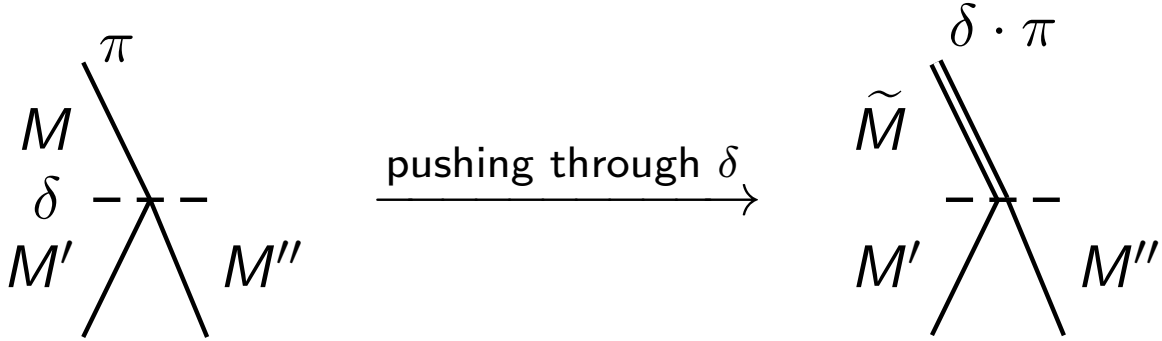
Convergence



Action of the splitting operator on an edge.

- $\widetilde{M}$ plays the role of implementing correlated changes.

Convergence



Action of the splitting operator on an edge.

- $\widetilde{M}$ plays the role of implementing correlated changes.
- Sumner et al. [2012] showed that the model can also be used for convergence.

Convergence

- Examples of convergence are hybridisation, horizontal gene transfer or convergence of morphological traits.

Convergence

- Examples of convergence are hybridisation, horizontal gene transfer or convergence of morphological traits.
- Compare non-clock-like and clock-like trees to networks with convergence.

Convergence

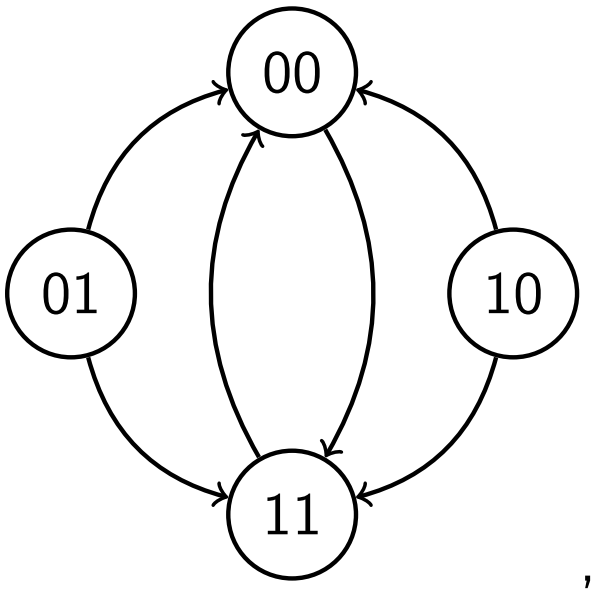
- Examples of convergence are hybridisation, horizontal gene transfer or convergence of morphological traits.
- Compare non-clock-like and clock-like trees to networks with convergence.
- Are our convergence-divergence networks **identifiable**?

Convergence

- Examples of convergence are hybridisation, horizontal gene transfer or convergence of morphological traits.
- Compare non-clock-like and clock-like trees to networks with convergence.
- Are our convergence-divergence networks **identifiable**?
- Can our convergence-divergence networks be **distinguished** from simpler trees?

Convergence-Divergence Network

Convergence of two taxa is as follows:

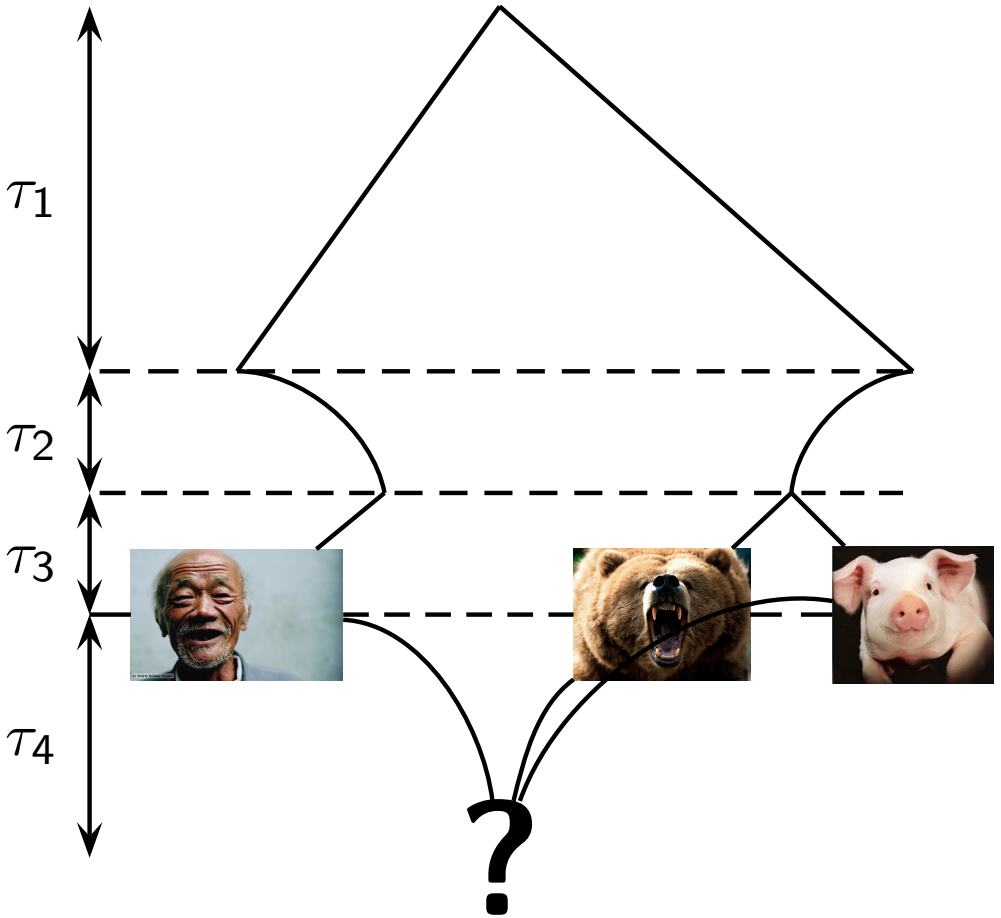


,

Q on two taxa.

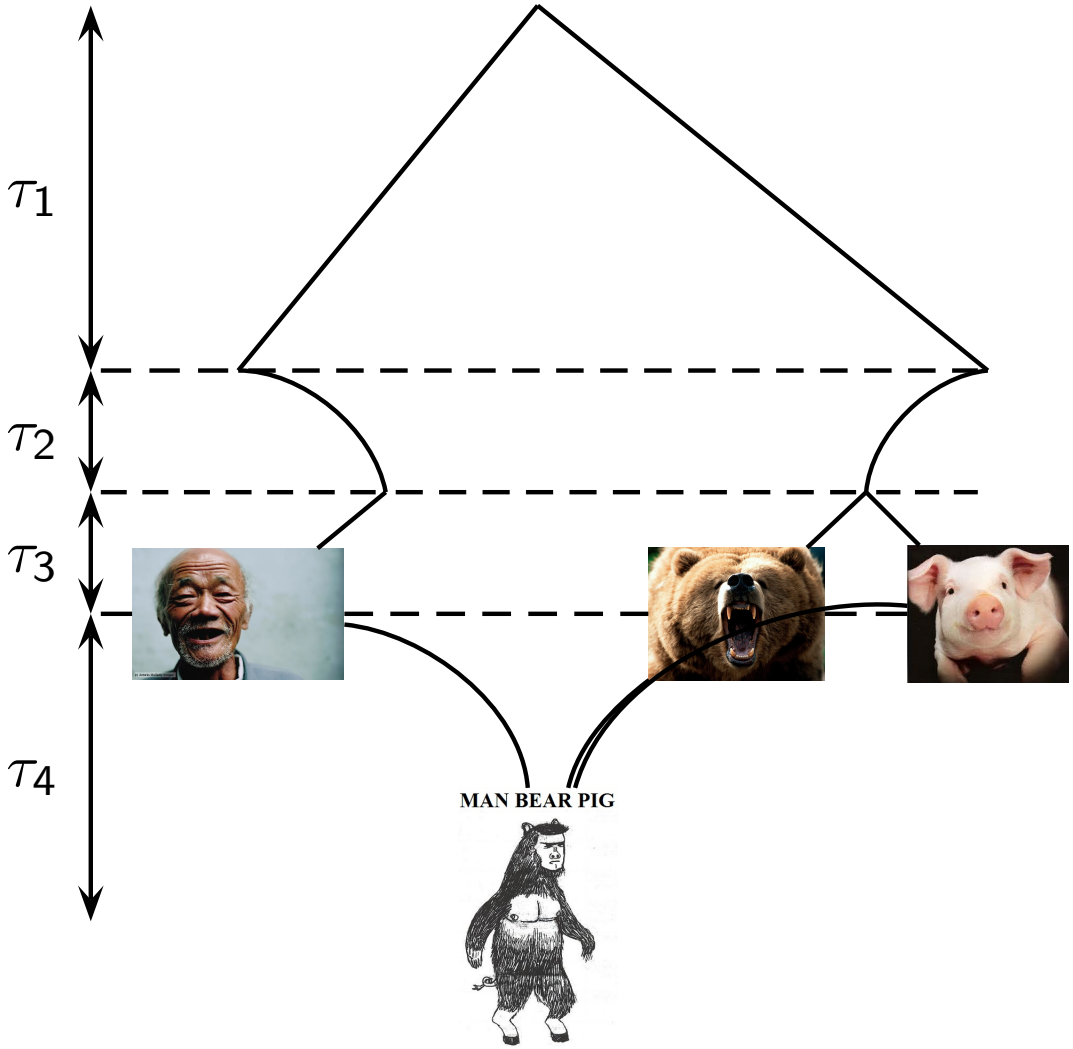
with each character state transition having the same rate, λ, from the binary symmetrical model.

Convergence-Divergence Network



A three-taxon convergence-divergence network.

Convergence-Divergence Network



A three-taxon convergence-divergence network.

Process

- Given a tree or network,

Process

- Given a tree or network,
  1. Transform the basis of the rate matrix of the model, eg. Hadamard transformation.

Process

- Given a tree or network,
  1. Transform the basis of the rate matrix of the model, eg. Hadamard transformation.
  2. Determine the probability distribution of the tree or network.

Process

- Given a tree or network,
  1. Transform the basis of the rate matrix of the model, eg. Hadamard transformation.
  2. Determine the probability distribution of the tree or network.
  3. Determine if the time parameters can be recovered from the probability distribution (**identifiability**).

Process

- Given a tree or network,
  1. Transform the basis of the rate matrix of the model, eg. Hadamard transformation.
  2. Determine the probability distribution of the tree or network.
  3. Determine if the time parameters can be recovered from the probability distribution (**identifiability**).
  4. Determine the constraints on the probability distribution, i.e. the probability space.

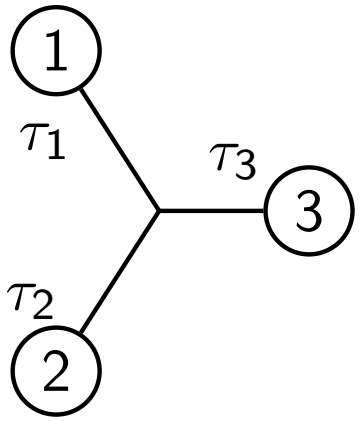
Process

- Given a tree or network,
  1. Transform the basis of the rate matrix of the model, eg. Hadamard transformation.
  2. Determine the probability distribution of the tree or network.
  3. Determine if the time parameters can be recovered from the probability distribution (**identifiability**).
  4. Determine the constraints on the probability distribution, i.e. the probability space.
- From here we can compare the probability spaces of competing trees and networks.

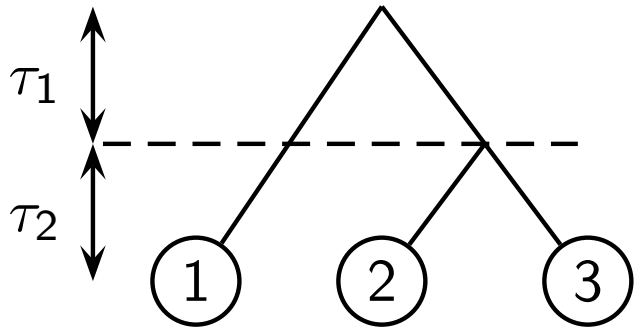
Process

- Given a tree or network,
  1. Transform the basis of the rate matrix of the model, eg. Hadamard transformation.
  2. Determine the probability distribution of the tree or network.
  3. Determine if the time parameters can be recovered from the probability distribution (**identifiability**).
  4. Determine the constraints on the probability distribution, i.e. the probability space.
- From here we can compare the probability spaces of competing trees and networks.
- If two trees or networks have the same probability space they are said to not be **distinguishable**.

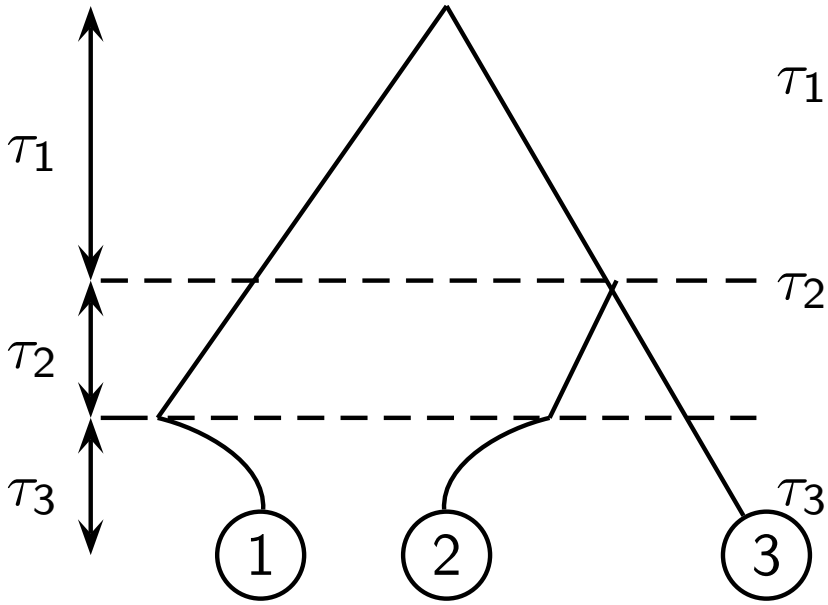
Three-Taxon Networks



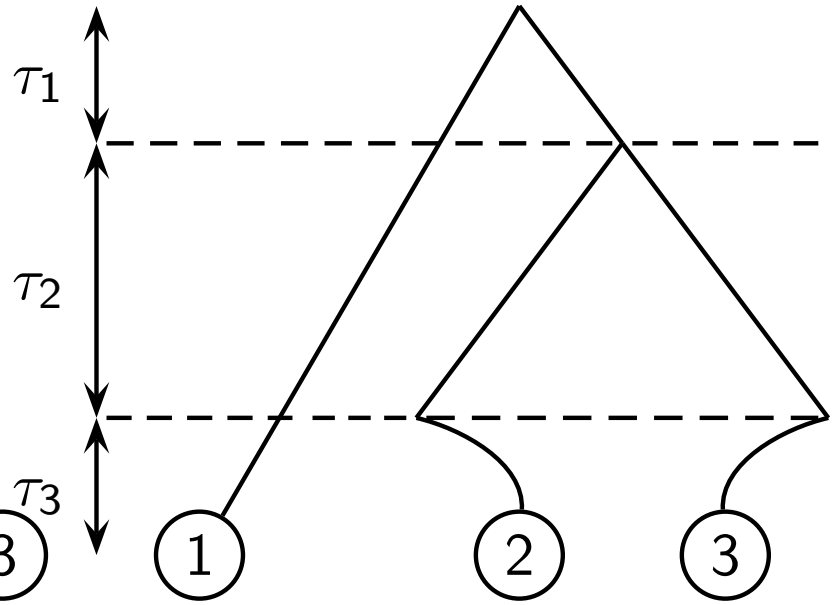
Network 1



Network 2

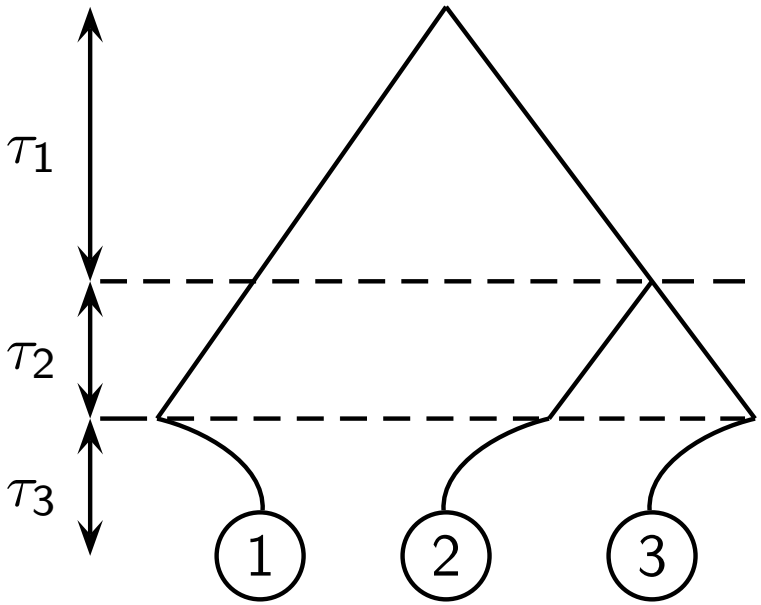


Network 3

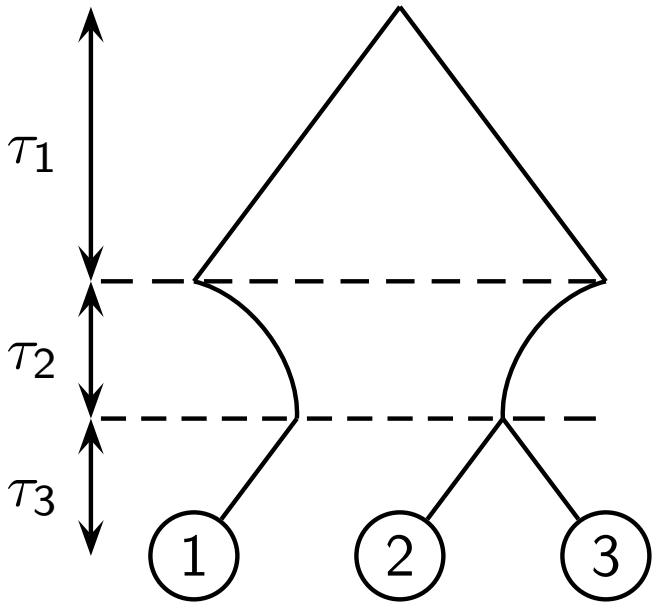


Network 4

Three-Taxon Networks

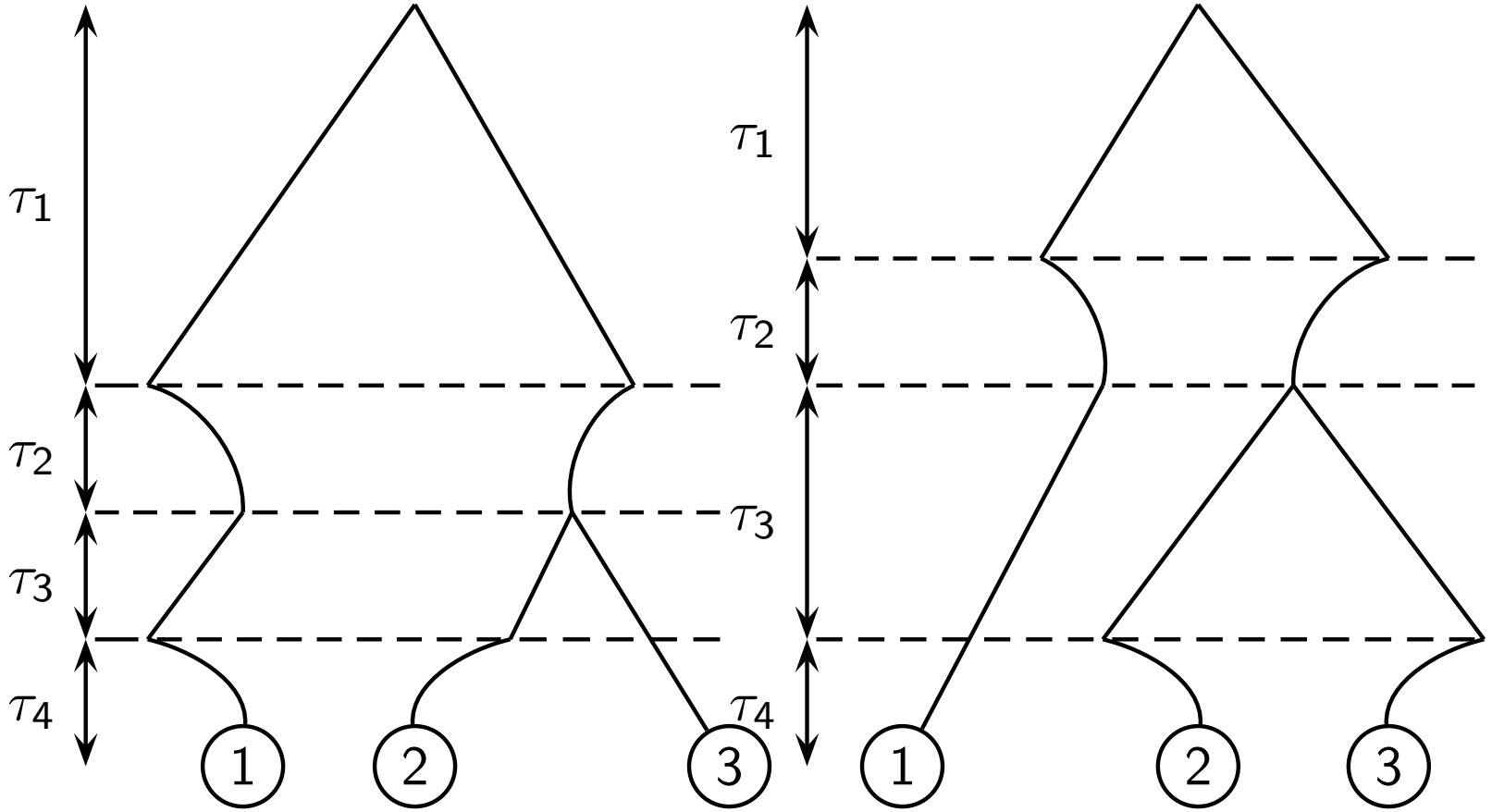


Network 5



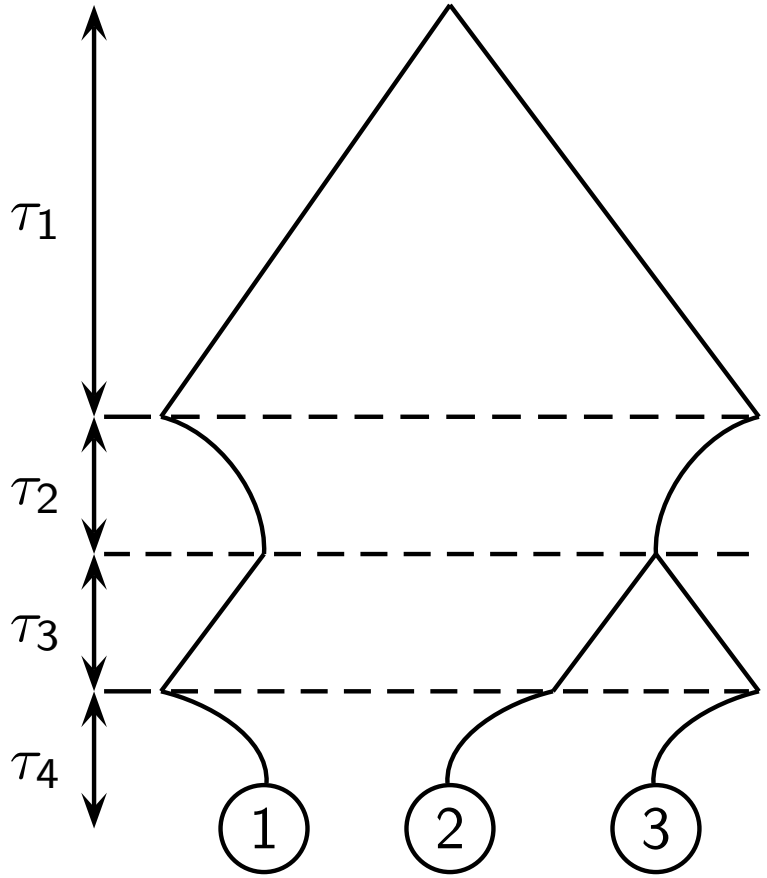
Network 6

Three-Taxon Networks



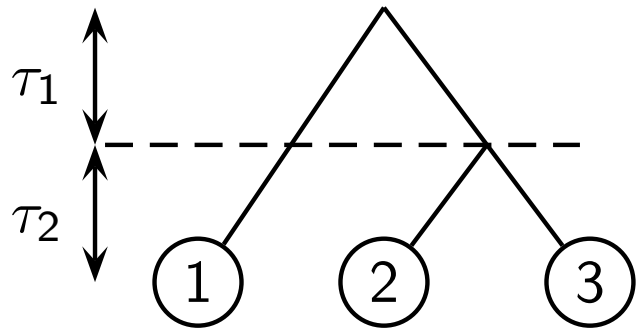
Network 7 Network 8

Three-Taxon Networks



Network 9

An Example: Three-Taxon Clock-Like Tree



Three-taxon clock-like tree.

- In the regular basis, P, and the Hadamard basis, $\widehat{P}$, the probability distribution is

$$P = \begin{bmatrix} p_{000} \\ p_{001} \\ p_{010} \\ p_{011} \\ p_{100} \\ p_{101} \\ p_{110} \\ p_{111} \end{bmatrix}, \quad \widehat{P} = \begin{bmatrix} q_{000} \\ q_{001} \\ q_{010} \\ q_{011} \\ q_{100} \\ q_{101} \\ q_{110} \\ q_{111} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ e^{-2\tau_2} \\ 0 \\ e^{-2(\tau_1+\tau_2)} \\ e^{-2(\tau_1+\tau_2)} \\ 0 \end{bmatrix}.$$

An Example: Three-Taxon Clock-Like Tree

- Make the substitutions, $x_i = e^{-\tau_i}$, to convert to polynomial functions.

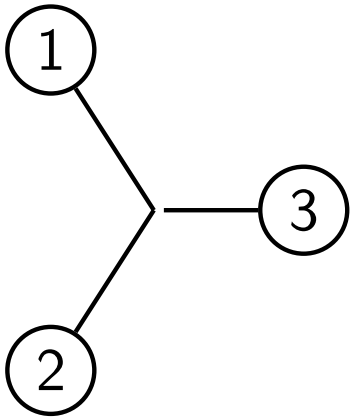
An Example: Three-Taxon Clock-Like Tree

- Make the substitutions, $x_i = e^{-\tau_i}$, to convert to polynomial functions.
- For the three-taxon clock-like tree,
  $\{q_{011} = x_2^2, \quad q_{101} = x_1^2 x_2^2, \quad q_{110} = x_1^2 x_2^2\}$.

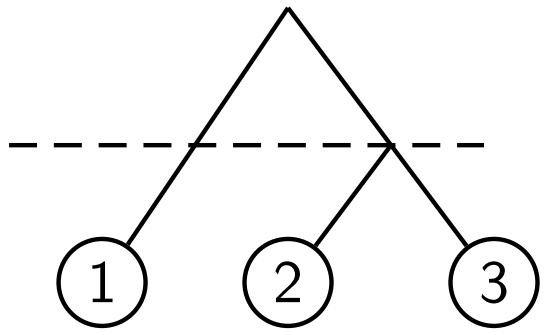
An Example: Three-Taxon Clock-Like Tree

- Make the substitutions, $x_i = e^{-\tau_i}$, to convert to polynomial functions.
- For the three-taxon clock-like tree,
  $\{q_{011} = x_2^2, \quad q_{101} = x_1^2 x_2^2, \quad q_{110} = x_1^2 x_2^2\}$.
- Constraints are, $\{q_{101} = q_{110}, \quad q_{011} \geq q_{101}\}$.

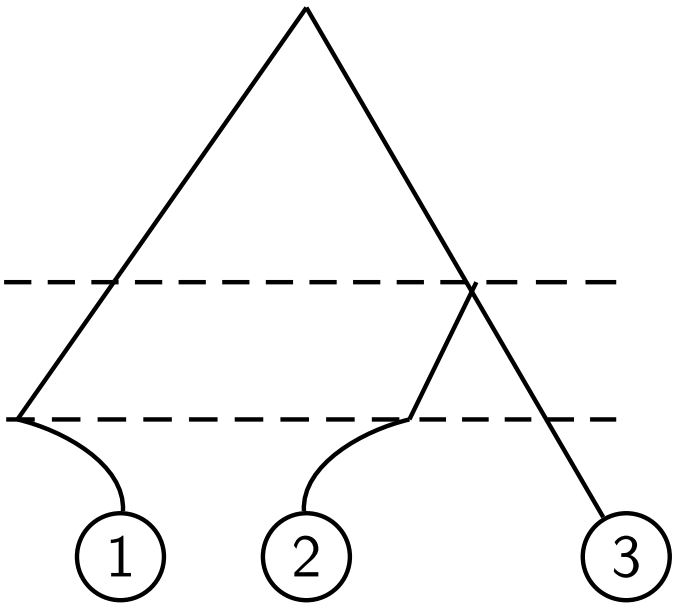
Three-Taxon Networks



Network 1



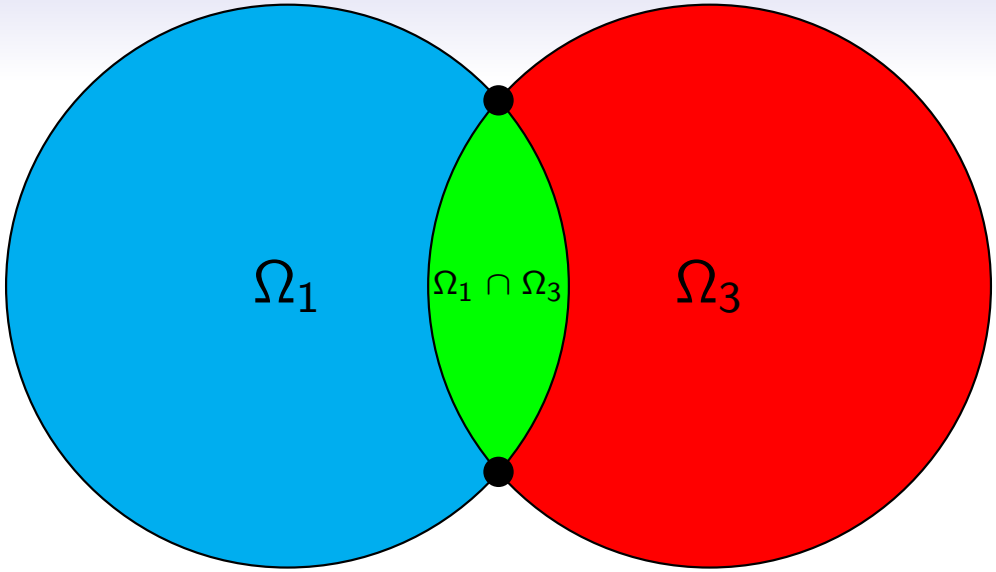
Network 2



Network 3

Network(s)	$q_{101} = q_{110}$ (Y/N)	$q_{110} \geq q_{101}$ (Y/N)	$q_{011} \geq q_{101}$ (Y/N)	$q_{011}(1-q_{110})^2 \geq (q_{011}-q_{101})^2$ (Y/N)
1	N	N	N	N
2, 4, 5, 6, 8, 9	Y	N	Y	N
3, 7	N	Y	Y	Y

In addition, the non-clock-like tree (Network 1) must meet the constraints $\{q_{011} \geq q_{101}q_{110}, \quad q_{101} \geq q_{011}q_{110}, \quad q_{110} \geq q_{011}q_{101}\}$.



Probability spaces of the networks. The probability space for Network 2 is the two black dots where the probability spaces for networks 1 and 3 intersect. Not to scale.

Colour	Probability Space	Constraints
Blue	Ω_1	$\{q_{011} \geq q_{101}q_{110}, \quad q_{101} \geq q_{011}q_{110}, \quad q_{110} \geq q_{011}q_{101}\}$
Red	Ω_3	$\{q_{011} \geq q_{101}, \quad q_{110} \geq q_{101}, \quad q_{011}(1-q_{110})^2 \geq (q_{011}-q_{101})^2\}$
Green	$\Omega_1 \cap \Omega_3$	$\{q_{011} \geq q_{101}, \quad q_{110} \geq q_{101}, \quad q_{101} \geq q_{011}q_{110}\}$
Black	$\Omega_1 \cap \Omega_2 \cap \Omega_3$	$\{q_{101} = q_{110}, \quad q_{011} \geq q_{110}\}$

Summary of network constraints which must be met in the region of the probability space.

- As an example, the constraints for the black region in regular basis are $\{p_{010} + p_{101} = p_{001} + p_{110}, \quad p_{011} + p_{100} \geq p_{001} + p_{110}\}$.

Data Analysis

- How does the convergence-divergence network perform in a likelihood scenario (BIC)?

Data Analysis

- How does the convergence-divergence network perform in a likelihood scenario (BIC)?
- Cormorants and Shags data set from Siegel-Causey [1988].

Data Analysis

- How does the convergence-divergence network perform in a likelihood scenario (BIC)?
- Cormorants and Shags data set from Siegel-Causey [1988].
- Binary morphological character sequence.

Data Analysis

- How does the convergence-divergence network perform in a likelihood scenario (BIC)?
- Cormorants and Shags data set from Siegel-Causey [1988].
- Binary morphological character sequence.
- Holland et al. [2010] showed that there appeared to be convergence of morphological traits.

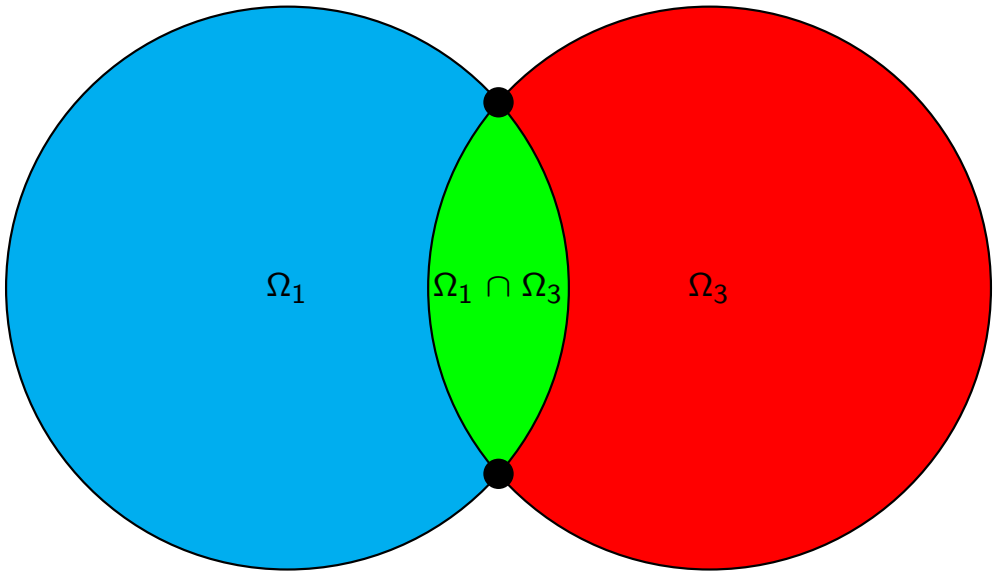
Data Analysis

- How does the convergence-divergence network perform in a likelihood scenario (BIC)?
- Cormorants and Shags data set from Siegel-Causey [1988].
- Binary morphological character sequence.
- Holland et al. [2010] showed that there appeared to be convergence of morphological traits.
- 30 taxa (not including 3 outgroups) and 137 sites, with some taxa missing data at particular sites.

Data Analysis

- How does the convergence-divergence network perform in a likelihood scenario (BIC)?
- Cormorants and Shags data set from Siegel-Causey [1988].
- Binary morphological character sequence.
- Holland et al. [2010] showed that there appeared to be convergence of morphological traits.
- 30 taxa (not including 3 outgroups) and 137 sites, with some taxa missing data at particular sites.
- Analysed all sets of triplets.

Data Analysis



Probability spaces of the networks. The probability space for Network 2 is the two black dots where the probability spaces for networks 1 and 3 intersect. Not to scale.

BIC_{nc}	BIC_{cl}	BIC_{cd0+}	BIC_{cd2+}	BIC_{cd6+}	BIC_{cd10+}
0.0367	0.8547	0.1451	0.03153	0.009113	0.007635

Summary statistics.

Note: $BIC_{nc} + BIC_{cl} + BIC_{cd0+} = 0.0367 + 0.8547 + 0.1451 = 1.0365 > 1$. For some triplets the BIC values tied between two or more of the trees and networks. In these circumstances each tree or network was counted as having the lowest BIC value.

Future Work

- Confirm code and analysis is correct.

Future Work

- Confirm code and analysis is correct.
- Analyse another data set.

Future Work

- Confirm code and analysis is correct.
- Analyse another data set.
- Analyse four-taxon case and extend beyond binary symmetric model.

Future Work

- Confirm code and analysis is correct.
- Analyse another data set.
- Analyse four-taxon case and extend beyond binary symmetric model.
- For more than three taxa we can use methods from alegbraic geometry (Gröbner bases).

References

B. R. Holland, H. G. Spencer, T. H. Worthy, and M. Kennedy. Identifying cliques of convergent characters: Concerted evolution in the cormorants and shags. *Systematic Biology*, 59(4): 433–445, 2010.

D. Siegel-Causey. Phylogeny of the phalacrocoracidae. *Condor*, pages 885–905, 1988.

J. G. Sumner, B. R. Holland, and P. D. Jarvis. The algebra of the general markov model on phylogenetic trees and networks. *Bulletin of Mathematical Biology*, 74:858–880, 2012.