

Statistical and Computational Methods for Complex Systems

George G Vega Yon

University of Southern California, Department of Preventive Medicine

Job Talk at Universidad Adolfo Ibáñez

December 10, 2020



Part 1: On the Prediction of Gene Functions Using Phylogenetic Trees

Part 2: Exponential Random Graph Models for Small Networks

Part 3: Current Research

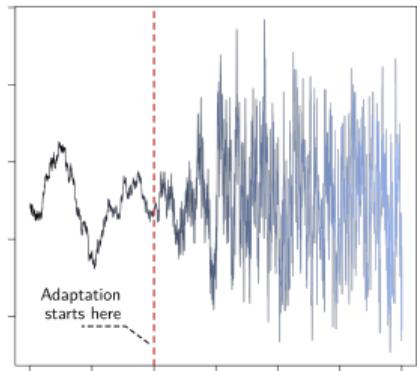
Part 4: Future Research

You can download the slides from <https://github.com/gvegayon/faculty-talk>

My work sits on the intersection between...

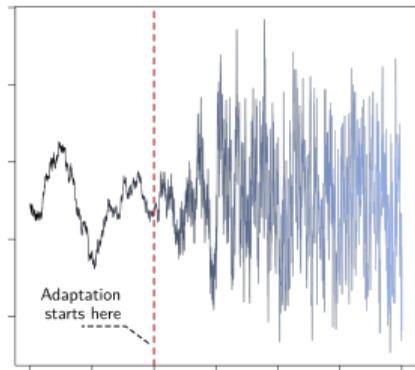
My work sits on the intersection between...

Statistics

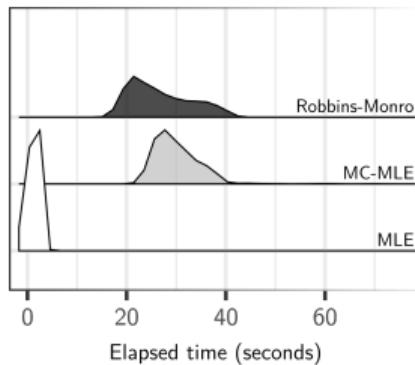


My work sits on the intersection between...

Statistics

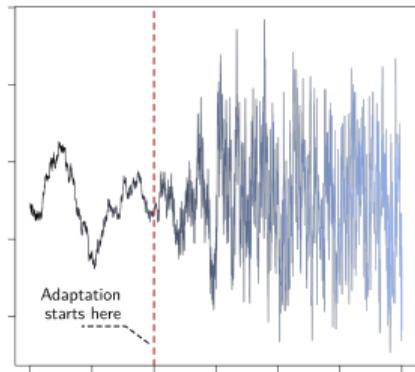


Computer Science

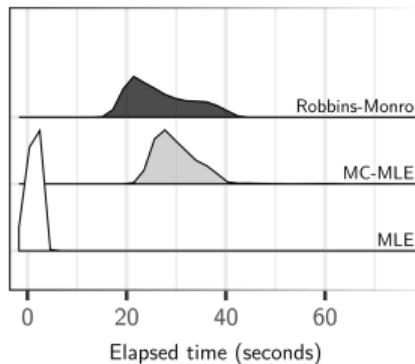


My work sits on the intersection between...

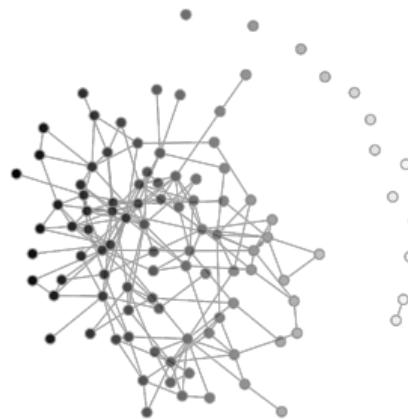
Statistics



Computer Science



Complex Systems

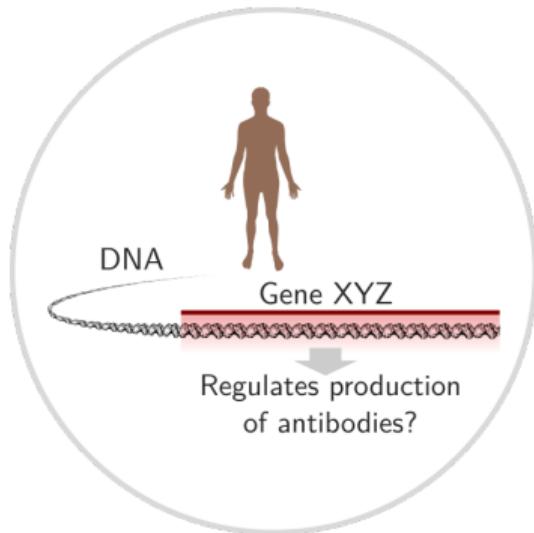


Part 1: On the Prediction of Gene Functions Using Phylogenetic Trees

Joint with: Paul D Thomas, Paul Marjoram, Huaiyu Mi, Duncan Thomas, and John Morrison
(2nd round of review Accepted at *PLOS Computational Biology*)

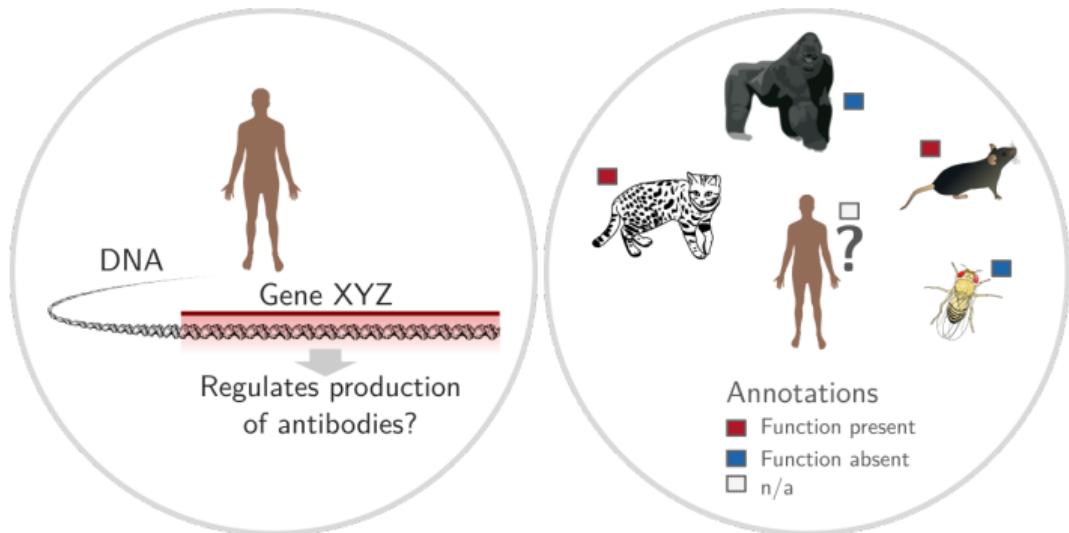
Uncovering the role of genes

Does Gene XYZ regulates antibody production?



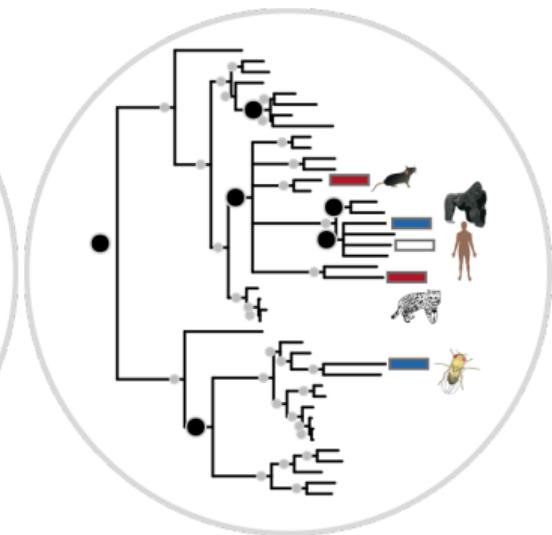
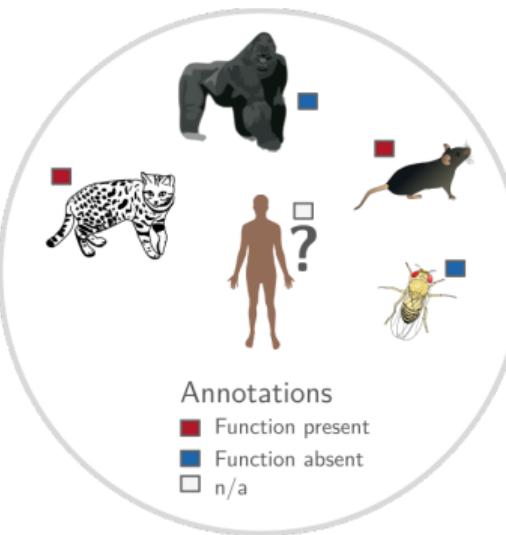
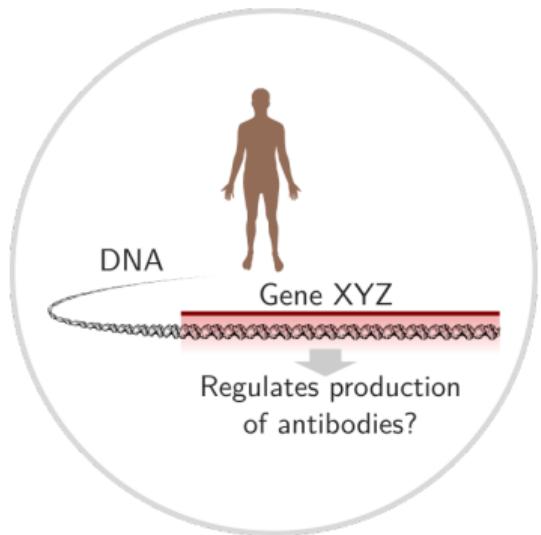
Uncovering the role of genes

Does Gene XYZ regulates antibody production?



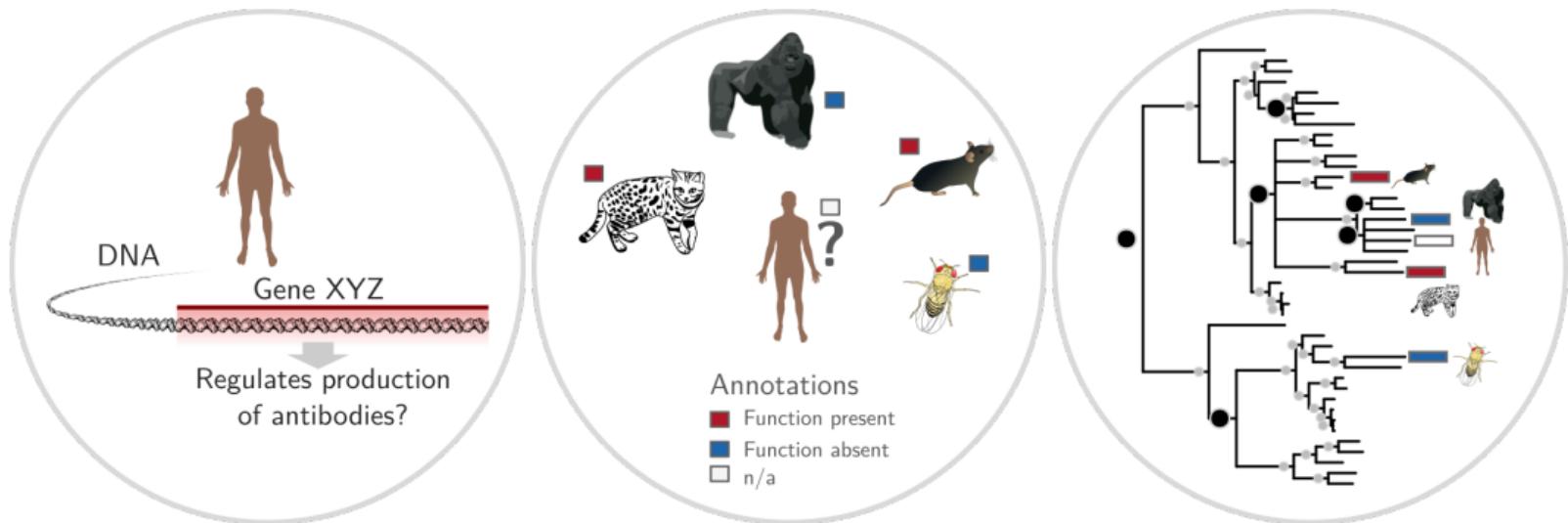
Uncovering the role of genes

Does Gene XYZ regulates antibody production?

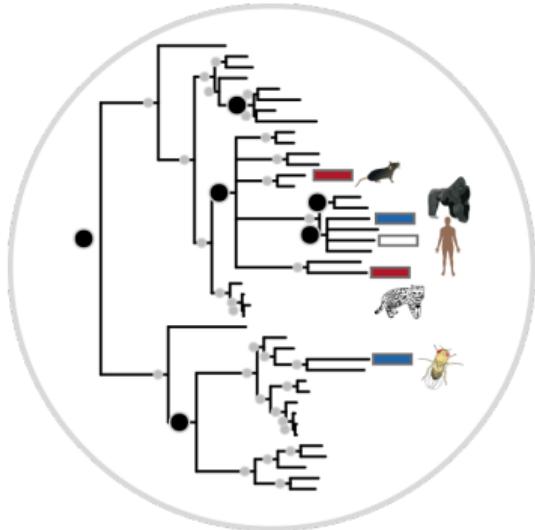


Uncovering the role of genes

Does Gene XYZ regulates antibody production?



... Phylogenetic trees can help us answer this question.



- ▶ ~ 15,000 evolutionary trees
- ▶ ~ 8 million annotations
- ▶ ~ 600 thousand on human genes
- ▶ ~ < 10% are based on experimental evidence

...Improving our knowledge on this is key to biomedical research

▶ more

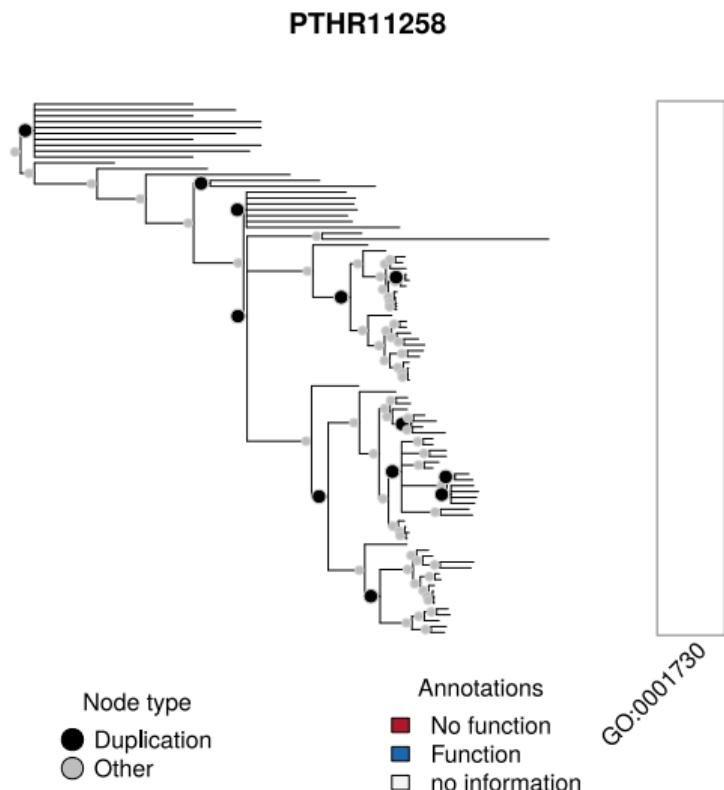
An evolutionary model of gene functions

Imagine a relay race...

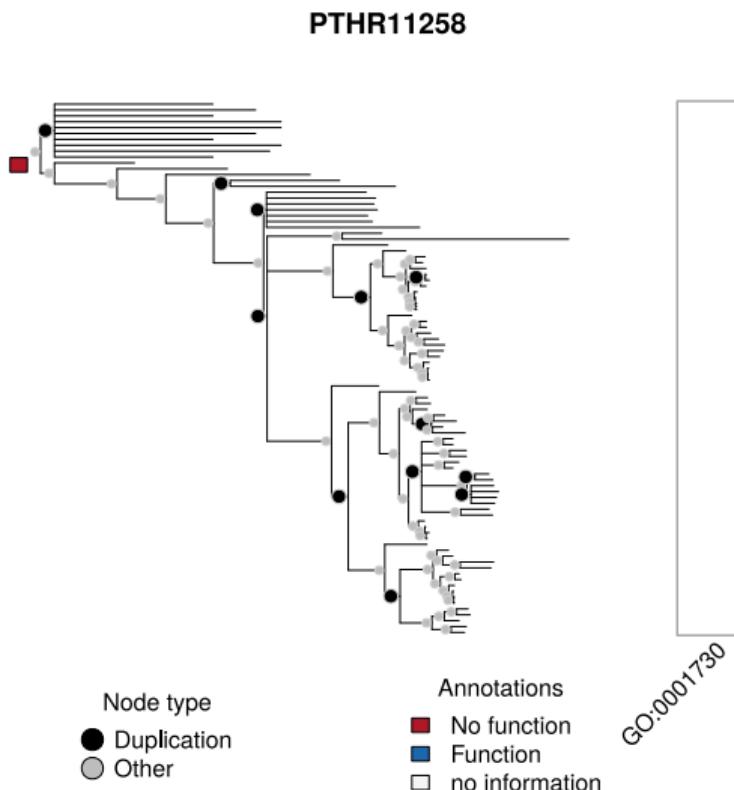


Figure 1 ISTAF 2019 4 × 100 m relay race (Martin Rulsch, wikimedia)

An evolutionary model of gene functions

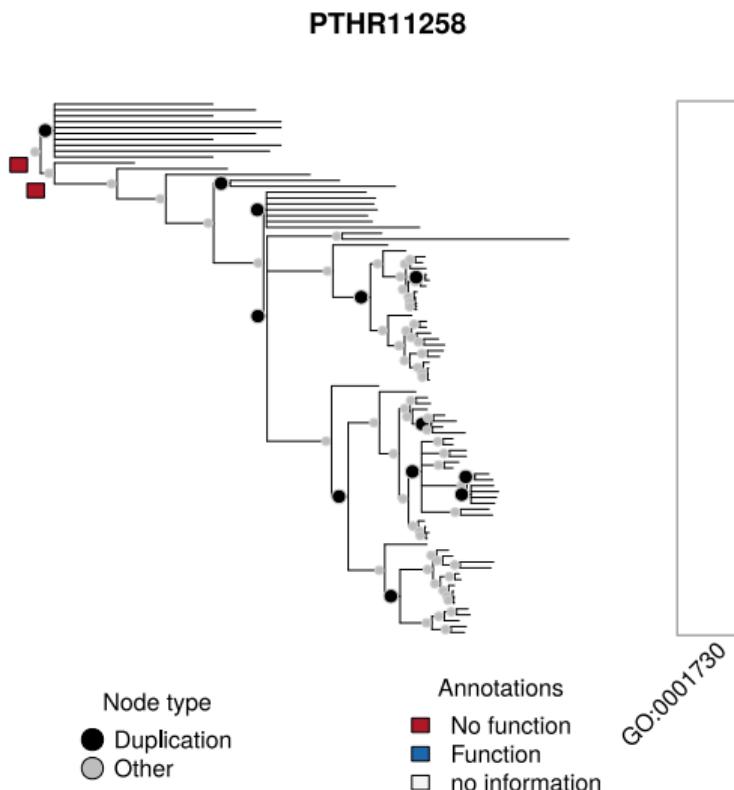


An evolutionary model of gene functions



- ▶ Starting with the root node (no function in this case).

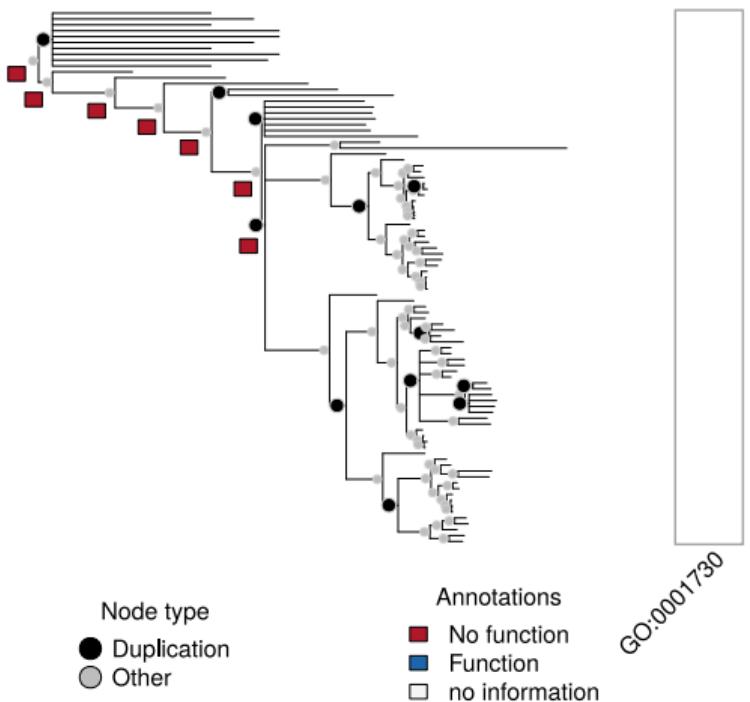
An evolutionary model of gene functions



- ▶ Starting with the root node (no function in this case).
- ▶ Passes the baton to its offspring.

An evolutionary model of gene functions

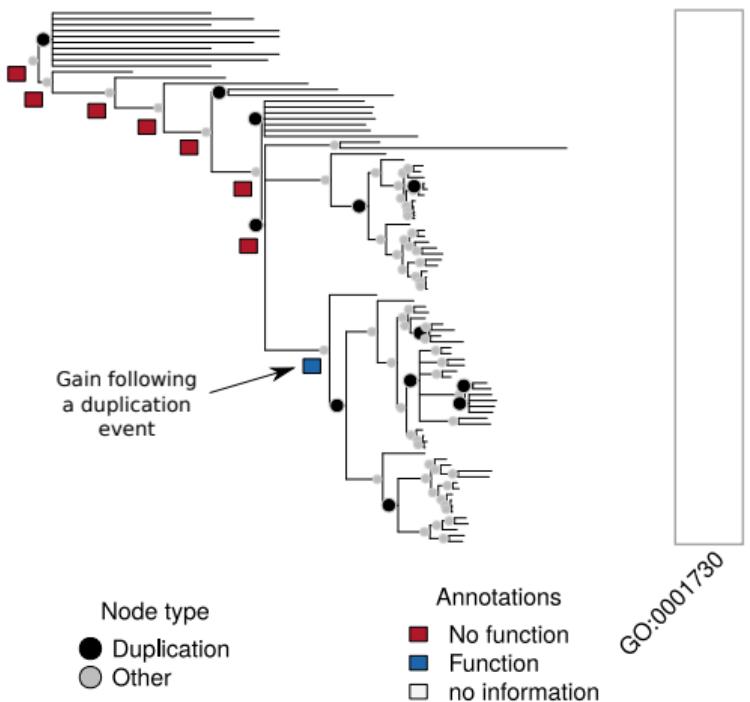
PTHR11258



- ▶ Starting with the root node (no function in this case).
- ▶ Passes the baton to its offspring.
- ▶ Possibly without change (on this particular function).

An evolutionary model of gene functions

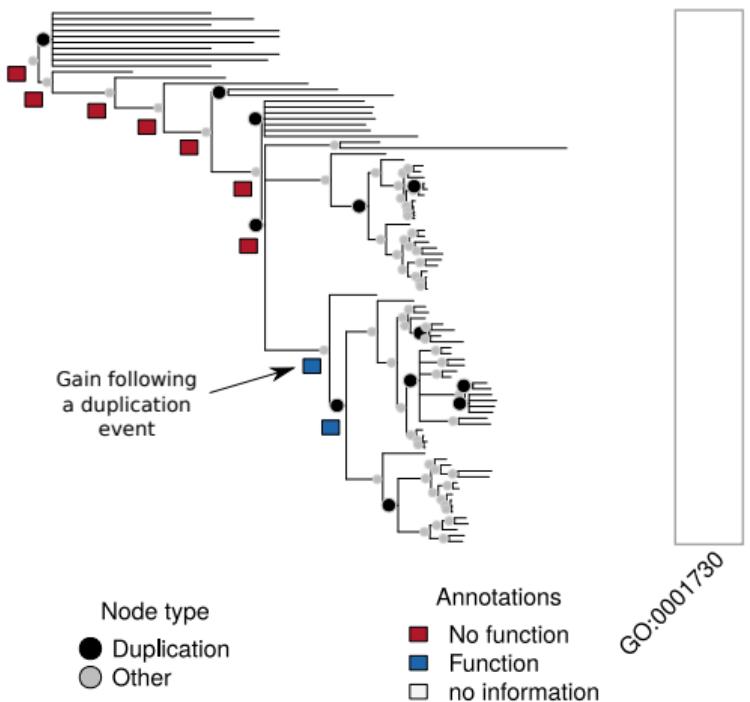
PTHR11258



- ▶ Starting with the root node (no function in this case).
- ▶ Passes the baton to its offspring.
- ▶ Possibly without change (on this particular function).
- ▶ Or, with some probability, gaining...

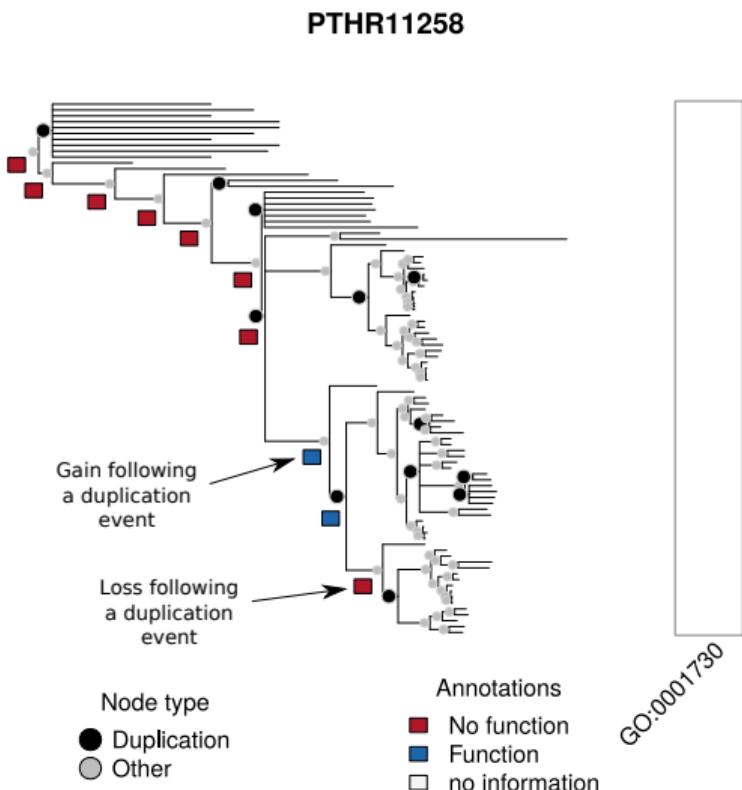
An evolutionary model of gene functions

PTHR11258



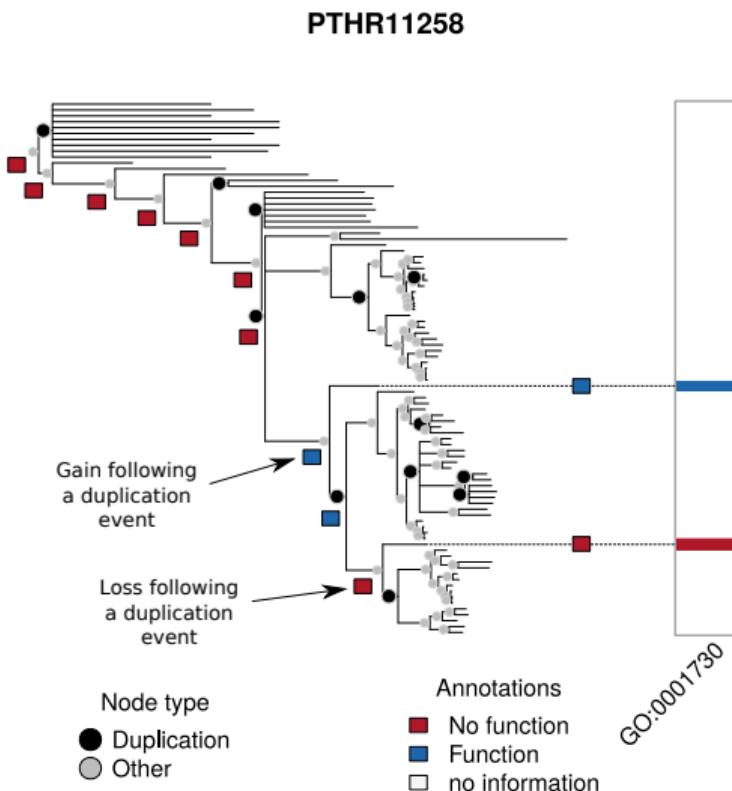
- ▶ Starting with the root node (no function in this case).
- ▶ Passes the baton to its offspring.
- ▶ Possibly without change (on this particular function).
- ▶ Or, with some probability, gaining...

An evolutionary model of gene functions



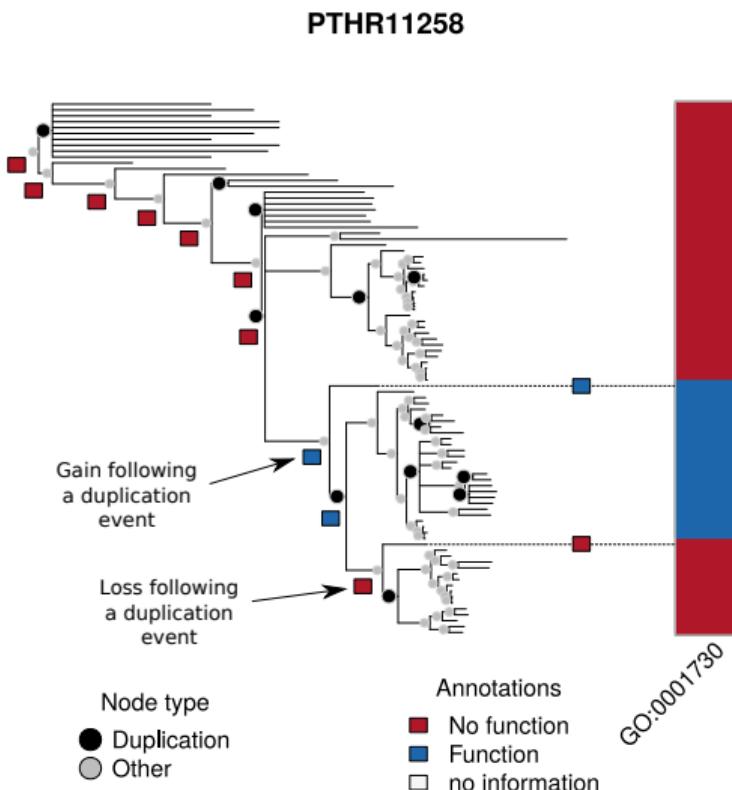
- ▶ Starting with the root node (no function in this case).
- ▶ Passes the baton to its offspring.
- ▶ Possibly without change (on this particular function).
- ▶ Or, with some probability, gaining... or loosing the function.

An evolutionary model of gene functions



- ▶ Starting with the root node (no function in this case).
- ▶ Passes the baton to its offspring.
- ▶ Possibly without change (on this particular function).
- ▶ Or, with some probability, gaining... or loosing the function.
- ▶ Until the *baton* reaches the end of the tree (modern genes).

An evolutionary model of gene functions



- ▶ Starting with the root node (no function in this case).
- ▶ Passes the baton to its offspring.
- ▶ Possibly without change (on this particular function).
- ▶ Or, with some probability, gaining... or loosing the function.
- ▶ Until the *baton* reaches the end of the tree (modern genes).

▶ more on duplication

▶ alt view

Example of Data + Predictions

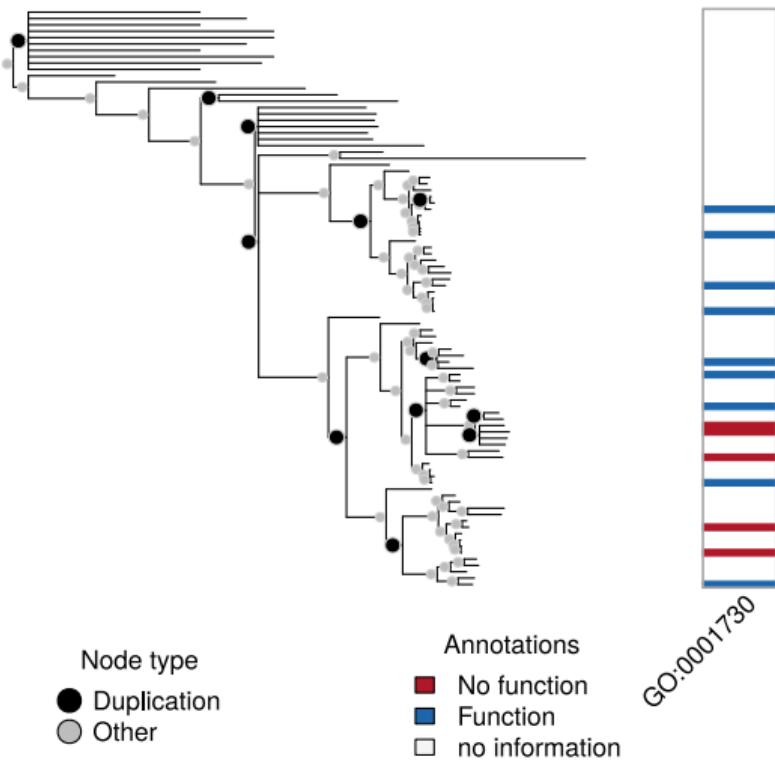
Family: PTHR11258

Type: Molecular Function

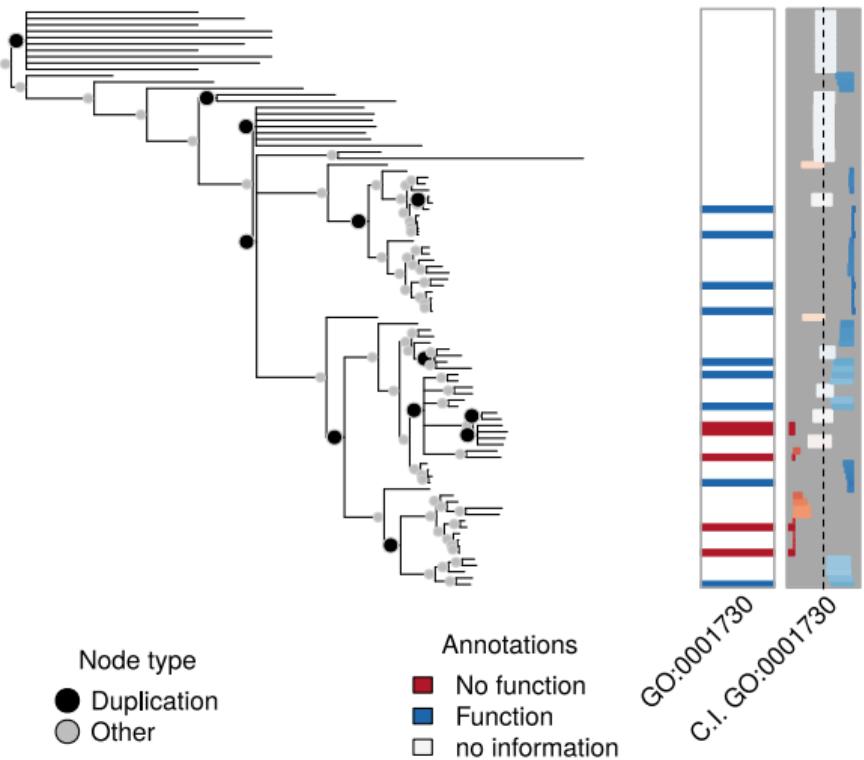
Name: 2'-5'-oligoadenylate synthetase activity

Desc: GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).

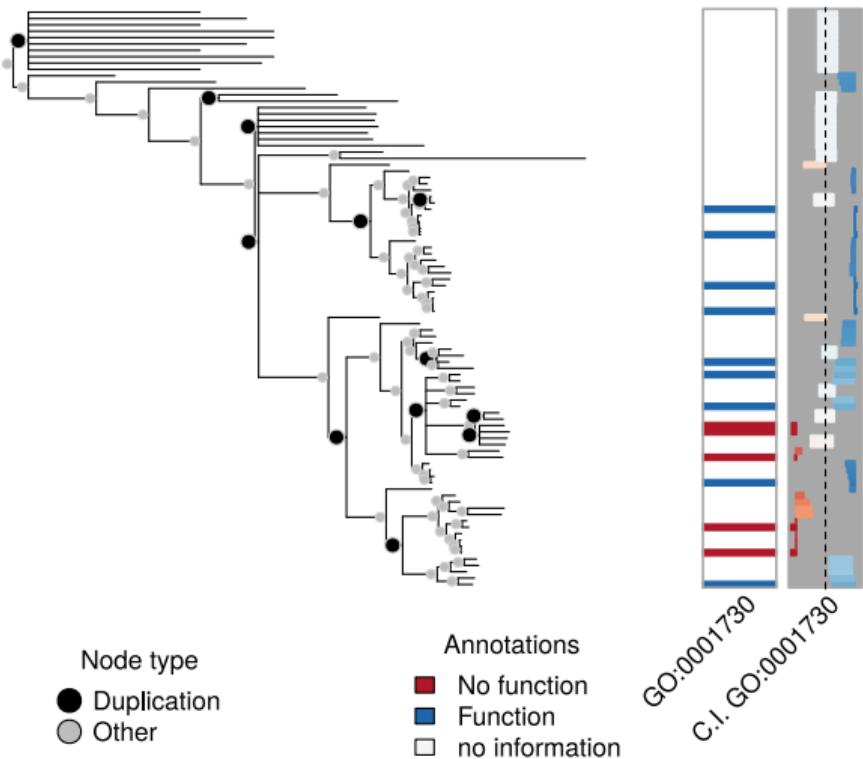
[see details](#)



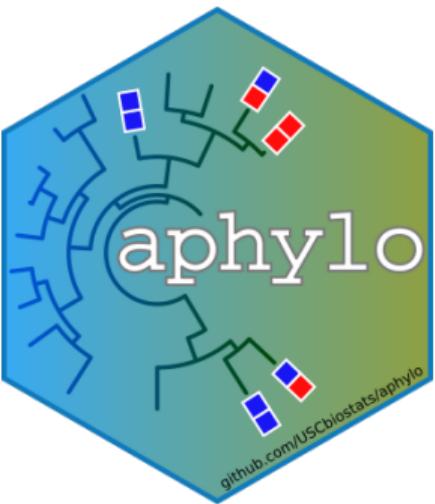
Example of Data + Predictions

Family: PTHR11258**Type:** Molecular Function**Name:** 2'-5'-oligoadenylate synthetase activity**Desc:** GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).[see details](#)

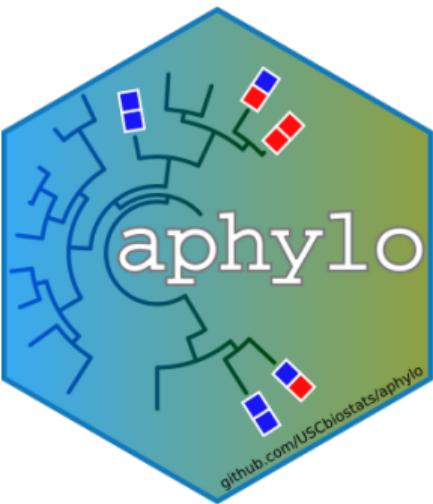
Example of Data + Predictions

Family: PTHR11258**Type:** Molecular Function**Name:** 2'-5'-oligoadenylate synthetase activity**Desc:** GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).**MAE:** 0.34**AUC:** 0.91[see details](#)

- ▶ Simulation, estimation, and prediction: **aphylo** R package.



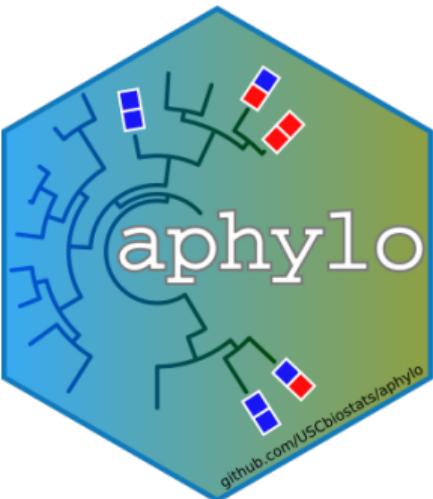
- ▶ Simulation, estimation, and prediction: **aphylo** R package.
- ▶ Large simulation study (all known trees, about 15,000) on USC's HPC cluster.



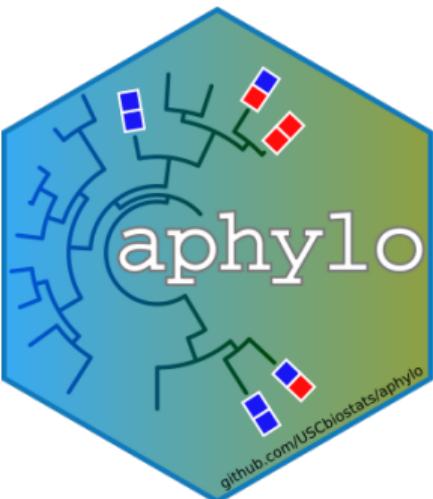
- ▶ Simulation, estimation, and prediction: **aphylo** R package.
- ▶ Large simulation study (all known trees, about 15,000) on USC's HPC cluster.
- ▶ Prediction quality assessment on $\sim 1,300$ genes involving ~ 130 families... estimation of parameters using a pooled-data model (< 5 min).

◀ modeling

◀ estimates



- ▶ Simulation, estimation, and prediction: **aphylo** R package.
- ▶ Large simulation study (all known trees, about 15,000) on USC's HPC cluster.
- ▶ Prediction quality assessment on $\sim 1,300$ genes involving ~ 130 families... estimation of parameters using a pooled-data model (< 5 min). [◀ modeling](#) [◀ estimates](#)
- ▶ In a subset of ~ 200 predictions we found 46 novel annotations [▶ more](#)



Results: Performance and Scalability

aphylo vs SIFTER (state-of-the-art phylo-based model) on 147 genes.

Results: Performance and Scalability

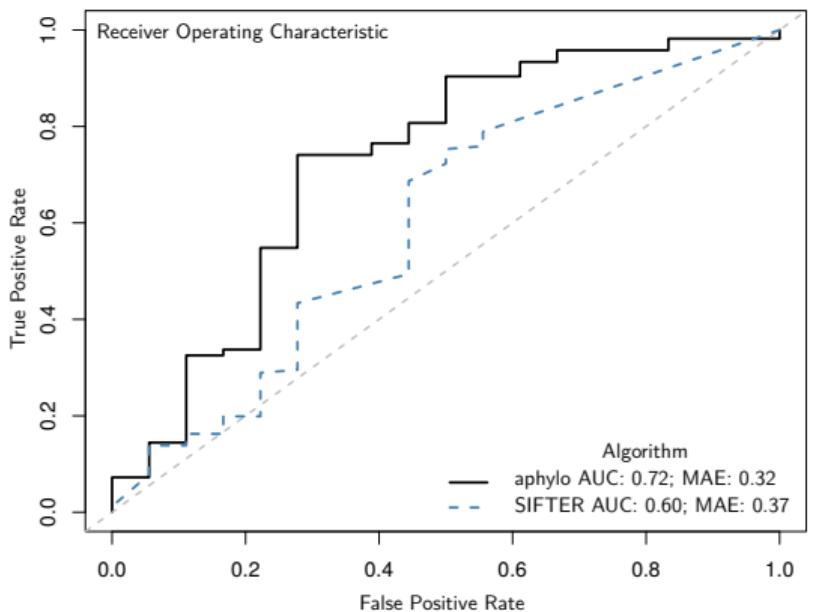
aphylo vs SIFTER (state-of-the-art phylo-based model) on 147 genes.



110 minutes (SIFTER) to calculate the posterior probabilities, aphylo took 1 second.

Results: Performance and Scalability

aphylo vs SIFTER (state-of-the-art phylo-based model) on 147 genes.



Fast 110 minutes (SIFTER) to calculate the posterior probabilities, aphylo took 1 second.

Accurate aphylo reported higher accuracy levels in LOO cross-validation (0.72 vs 0.60 AUC).

Computational Complexity: aphylo

Baseline features

Computational Complexity: aphylo

Baseline features

- ▶ Parsimony: Independence across functions/siblings.

Computational Complexity: aphylo

Baseline features

- ▶ Parsimony: Independence across functions/siblings.
- ▶ Post-order Tree traversal: Linear complexity $O(|\text{tree}|)$.

Computational Complexity: aphylo

Baseline features

- ▶ Parsimony: Independence across functions/siblings.
- ▶ Post-order Tree traversal: Linear complexity $O(|\text{tree}|)$.

Additional features

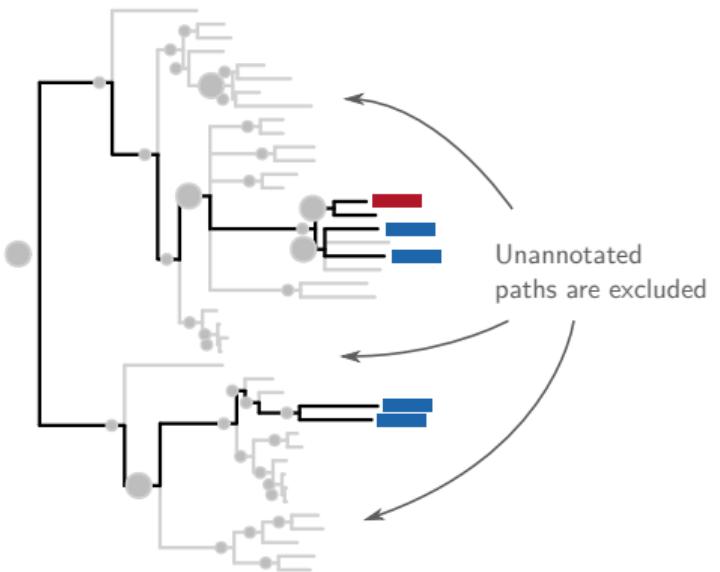
Computational Complexity: aphylo

Baseline features

- ▶ Parsimony: Independence across functions/siblings.
- ▶ Post-order Tree traversal: Linear complexity $O(|\text{tree}|)$.

Additional features

- ▶ Reduced pruning sequence: Induced sub-tree of nodes connected to annotated leafs
 \implies Complexity $O(|\text{Induced sub-tree}|) \leq O(|\text{tree}|)$



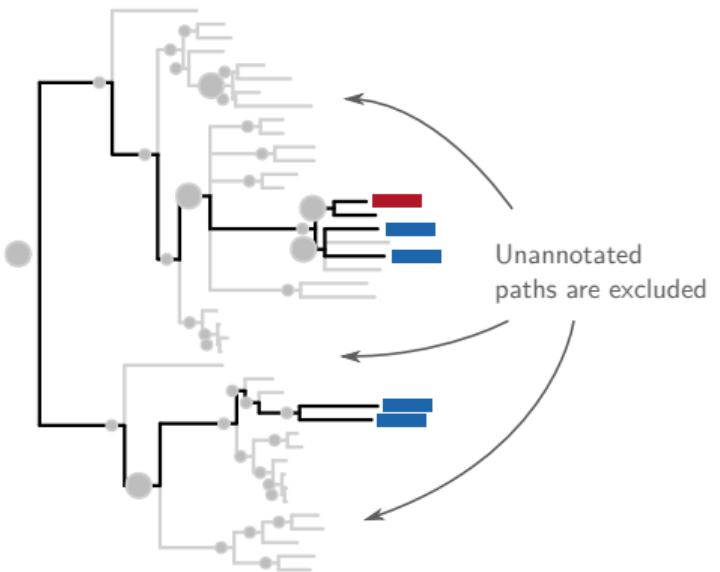
Computational Complexity: aphylo

Baseline features

- ▶ Parsimony: Independence across functions/siblings.
- ▶ Post-order Tree traversal: Linear complexity $O(|\text{tree}|)$.

Additional features

- ▶ Reduced pruning sequence: Induced sub-tree of nodes connected to annotated leafs
 \implies Complexity $O(|\text{Induced sub-tree}|) \leq O(|\text{tree}|)$
- ▶ Implemented in C++ (**pruner** library):



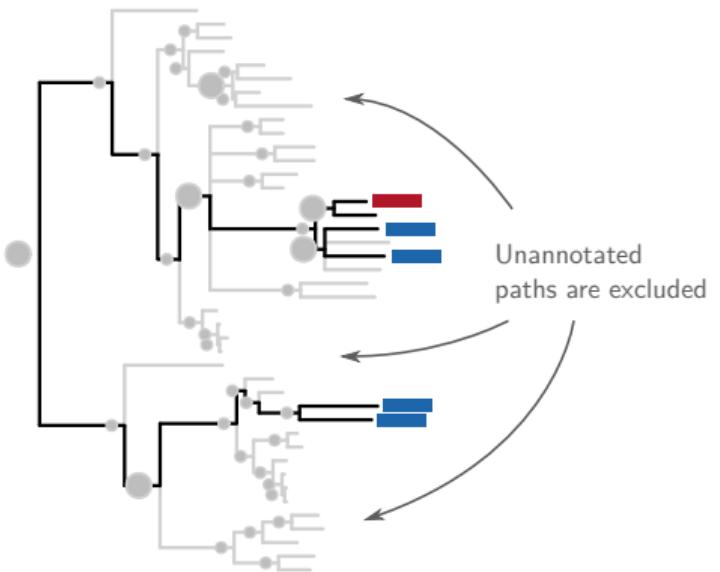
Computational Complexity: aphylo

Baseline features

- ▶ Parsimony: Independence across functions/siblings.
- ▶ Post-order Tree traversal: Linear complexity $O(|\text{tree}|)$.

Additional features

- ▶ Reduced pruning sequence: Induced sub-tree of nodes connected to annotated leafs
 - ⇒ Complexity $O(|\text{Induced sub-tree}|) \leq O(|\text{tree}|)$
- ▶ Implemented in C++ (**pruner** library):
 - ⇒ Fast



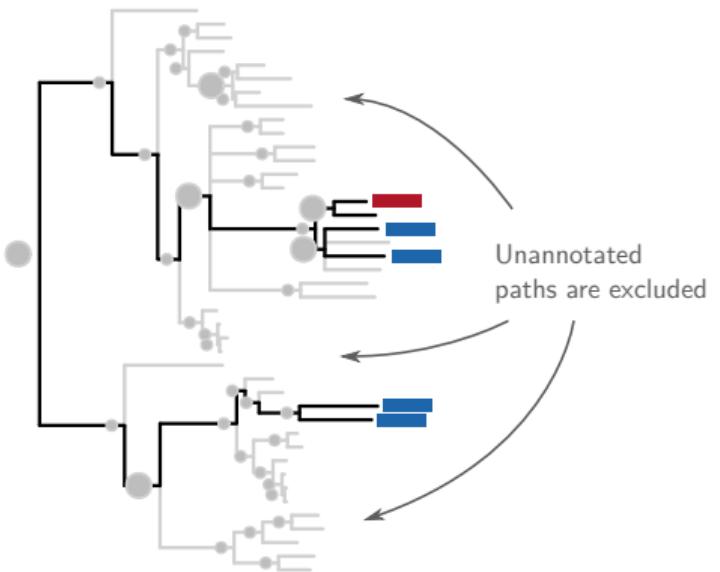
Computational Complexity: aphylo

Baseline features

- ▶ Parsimony: Independence across functions/siblings.
- ▶ Post-order Tree traversal: Linear complexity $O(|\text{tree}|)$.

Additional features

- ▶ Reduced pruning sequence: Induced sub-tree of nodes connected to annotated leafs
 - ⇒ Complexity $O(|\text{Induced sub-tree}|) \leq O(|\text{tree}|)$
- ▶ Implemented in C++ (**pruner** library):
 - ⇒ Fast
 - ⇒ Efficient use of memory (data structures)



Part 2: Exponential Random Graph Models for Small Networks

Joint with: Andrew Slaughter and Kayla de la Haye
(published in the journal *Social Networks*)



Data: Friendship network of a UK university faculty

from `igraphdata`. Viz: R package `netplot` (yours truly,
github.com/usccana/netplot)

- ▶ If COVID-19 has taught us something it is that networks matter.



Data: Friendship network of a UK university faculty
from `igraphdata`. Viz: R package `netplot` (yours truly,
github.com/usccana/netplot)

- ▶ If COVID-19 has taught us something it is that networks matter.
- ▶ And especially small networks: Families, teams, friends, etc.



Data: Friendship network of a UK university faculty
from `igraphdata`. Viz: R package `netplot` (yours truly,
github.com/usccana/netplot)

- ▶ If COVID-19 has taught us something it is that networks matter.
- ▶ And especially small networks: Families, teams, friends, etc. The cornerstone of larger social systems.



Data: Friendship network of a UK university faculty
from `igraphdata`. Viz: R package `netplot` (yours truly,
github.com/usccana/netplot)

- ▶ If COVID-19 has taught us something it is that networks matter.
- ▶ And especially small networks: Families, teams, friends, etc. The cornerstone of larger social systems.
- ▶ We can study networks using ERGMs.

Data: Friendship network of a UK university faculty
from `igraphdata`. Viz: R package `netplot` (yours truly,
github.com/usccana/netplot)



What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka **ERGMs** are:

What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka **ERGMs** are:

- ▶ Statistical models of (social) networks.

What are Exponential Random Graph Models

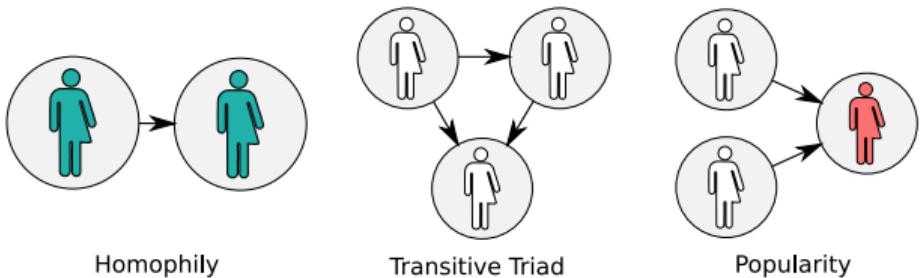
Exponential Family Random Graph Models, aka **ERGMs** are:

- ▶ Statistical models of (social) networks.
- ▶ Not about individual ties, but about local structures (sufficient statistics).

What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka **ERGMs** are:

- ▶ Statistical models of (social) networks.
- ▶ Not about individual ties, but about local structures (sufficient statistics).



Discrete Exponential-Family Models

$$\mathbb{P}(\mathbf{G} = \mathbf{g} \mid X = x) = \frac{\exp\{\boldsymbol{\theta}^t s(\mathbf{g}, x)\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \exp\{\boldsymbol{\theta}^t s(\mathbf{g}', x)\}}, \quad \forall \mathbf{g} \in \mathcal{G}$$

A vector of model parameters A vector of sufficient statistics

Observed data The normalizing constant All possible networks

A vector of model parameters	A vector of sufficient statistics
$\mathbb{P}(\mathbf{G} = \mathbf{g} \mid X = x) = \frac{\exp\{\theta^t s(\mathbf{g}, x)\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \exp\{\theta^t s(\mathbf{g}', x)\}}, \quad \forall \mathbf{g} \in \mathcal{G}$	
Observed data	The normalizing constant
	All possible networks

- For any 0/1 matrix of size $(n \times m)$, there are $2^{(n \times m)}$ possible realizations.

Discrete Exponential-Family Models

$$\mathbb{P}(\mathbf{G} = \mathbf{g} \mid X = x) = \frac{\exp\{\theta^t s(\mathbf{g}, x)\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \exp\{\theta^t s(\mathbf{g}', x)\}}, \quad \forall \mathbf{g} \in \mathcal{G}$$

A vector of model parameters A vector of sufficient statistics

 Observed data The normalizing constant All possible networks

- ▶ For any 0/1 matrix of size $(n \times m)$, there are $2^{(n \times m)}$ possible realizations.
 - ▶ A directed graph of size 5 has 1,048,576 possible configurations!

Discrete Exponential-Family Models

$$\mathbb{P}(\mathbf{G} = \mathbf{g} \mid X = x) = \frac{\exp\{\theta^t s(\mathbf{g}, x)\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \exp\{\theta^t s(\mathbf{g}', x)\}}, \quad \forall \mathbf{g} \in \mathcal{G}$$

A vector of model parameters A vector of sufficient statistics

 Observed data The normalizing constant All possible networks

- ▶ For any 0/1 matrix of size $(n \times m)$, there are $2^{(n \times m)}$ possible realizations.
 - ▶ A directed graph of size 5 has 1,048,576 possible configurations!
 - ▶ Most (all) applications use **approximations**...

A vector of model parameters	A vector of sufficient statistics	
θ	$s(\mathbf{g}, x)$	
		$\mathbb{P}(\mathbf{G} = \mathbf{g} \mid X = x) = \frac{\exp\{\theta^t s(\mathbf{g}, x)\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \exp\{\theta^t s(\mathbf{g}', x)\}}, \quad \forall \mathbf{g} \in \mathcal{G}$
Observed data	The normalizing constant	All possible networks

- ▶ For any 0/1 matrix of size $(n \times m)$, there are $2^{(n \times m)}$ possible realizations.
 - ▶ A directed graph of size 5 has 1,048,576 possible configurations!
 - ▶ Most (all) applications use **approximations**... yet, for sufficiently small graphs we “can be exact.”

► more theory

► more terms

Using **exact statistics** means:

- Higher convergence rate.
- Lower type-I error rate.
- Smaller bias.
- More power.
- Faster.
- More flexible.

▶ more

ergmito: Estimation of Small ERGM using Exact Statistics

Using exact statistics means:

- Higher convergence rate.
- Lower type-I error rate.
- Smaller bias.
- More power.
- Faster.
- More flexible.

CRAN - Package ergmito

cran.r-project.org/web/p... ergmito 0.2-1-9999

ergmito: Exponential Random Graph Models for Small Networks

Simulation and estimation of Exponential Random Graph Models (ERGMs) for small networks using exact statistics. As a difference from the 'ergm' package, 'ergmito' circumvents using Markov-Chain Maximum Likelihood Estimator (MC-MLE) and instead uses Maximum Likelihood Estimator (MLE) to fit ERGMs for small networks. As exhaustive enumeration is computationally feasible for small networks, this R package takes advantage of this and provides tools for calculating likelihood functions, and other relevant functions, directly, meaning that in many cases both estimation and simulation of ERGMs for small networks can be faster and more accurate than simulation-based algorithms.

Version: 0.2-1
Depends: R (≥ 3.3.0)
Imports: ergm, network, MASS, Rcpp, texreg, stats, parallel, utils, methods, graphics
LinkingTo: Rcpp, RcppArmadillo
Suggests: covr, sna, lmtest, fmcmc, coda, knitr, rmarkdown, tinytest
Published: 2020-02-12
Author: George Vega Yon [cre, aut], Kayla de la Haye [ths], Army Research Laboratory and the U.S. Army Research Office [fnd] (Grant Number W911NF-15-1-0577)

Exponential Random Graph

muriteams.github.io/ergmito/ ergmito 0.2-1-9999

ergmito: Exponential Random Graph Models for Small Networks



This R package, which has been developed on top of the amazing work that the Statnet team has done, implements estimation and simulation methods for Exponential Random Graph Models of small networks, in particular, less than 7 nodes. In the case of small networks, the calculation of the likelihood of ERGMs becomes computationally feasible, which allows us avoiding approximations and do exact calculation, ultimately obtaining MLEs directly.

Support

This material is based upon work support by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under grant number W911NF-15-1-0577

Computation for the work described in this paper was supported by the University of Southern California's Center for High-Performance Computing (www.usc.edu/cmp).

more

ergmito: Estimation of Small ERGM using Exact Statistics

Using exact statistics means:

- Higher convergence rate.
- Lower type-I error rate.
- Smaller bias.
- More power.
- Faster.
- More flexible.

downloads 5649

CRAN - Package ergmito

ergmito: Exponential Random Graph Models for Small Networks

Simulation and estimation of Exponential Random Graph Models (ERGMs) for small networks using exact statistics. As a difference from the 'ergm' package, 'ergmito' circumvents using Markov-Chain Maximum Likelihood Estimator (MC-MLE) and instead uses Maximum Likelihood Estimator (MLE) to fit ERGMs for small networks. As exhaustive enumeration is computationally feasible for small networks, this R package takes advantage of this and provides tools for calculating likelihood functions, and other relevant functions, directly, meaning that in many cases both estimation and simulation of ERGMs for small networks can be faster and more accurate than simulation-based algorithms.

Version: 0.2-1
Depends: R (≥ 3.3.0)
Imports: ergm, network, MASS, Rcpp, texreg, stats, parallel, utils, methods, graphics
LinkingTo: Rcpp, RcppArmadillo
Suggests: covr, sna, lmtest, fmcmc, coda, knitr, rmarkdown, tinytest
Published: 2020-02-12
Author: George Vega Yon [cre, aut], Kayla de la Haye [ths], Army Research Laboratory and the U.S. Army Research Office [fnd] (Grant Number W911NF-15-1-0577)

Exponential Random Graph

ergmito 0.2-1-9999

ergmito: Exponential Random Graph Models for Small Networks



This R package, which has been developed on top of the amazing work that the Statnet team has done, implements estimation and simulation methods for Exponential Random Graph Models of small networks, in particular, less than 7 nodes. In the case of small networks, the calculation of the likelihood of ERGMs becomes computationally feasible, which allows us avoiding approximations and do exact calculation, ultimately obtaining MLEs directly.

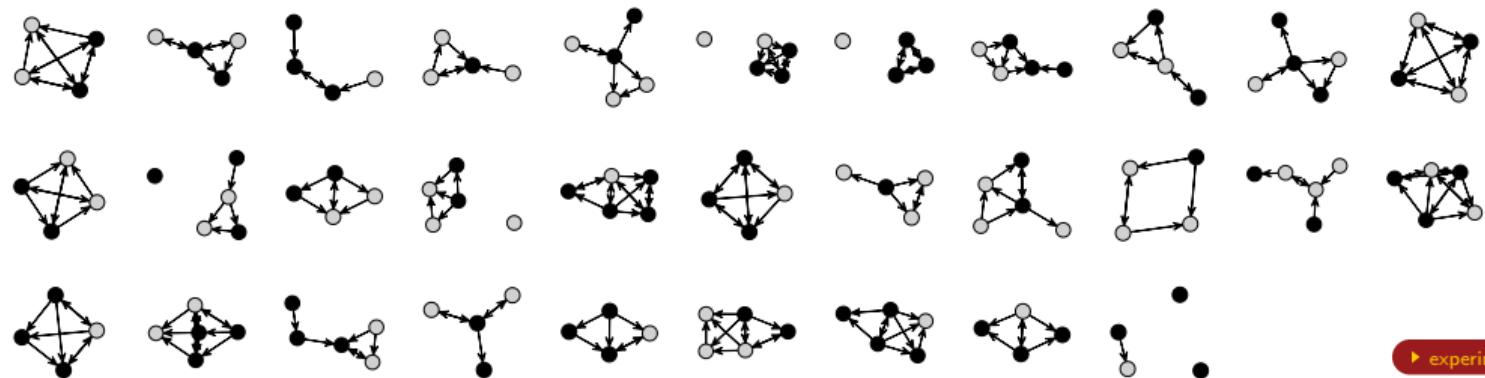
Support

This material is based upon work support by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under grant number W911NF-15-1-0577

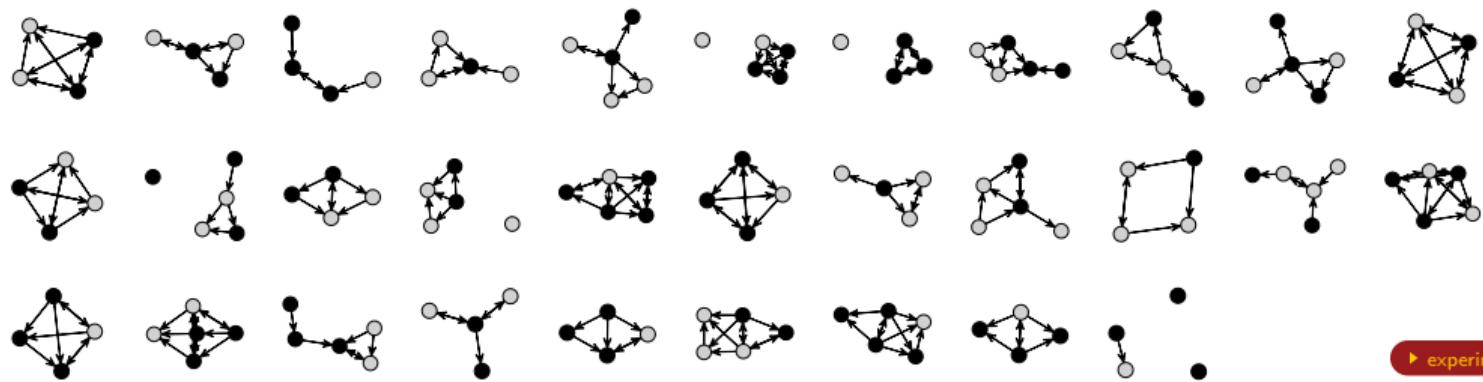
Computation for the work described in this paper was supported by the University of Southern California's Center for High-Performance Computing (www.usc.edu/cmp).

more

ergmito: Featured Example

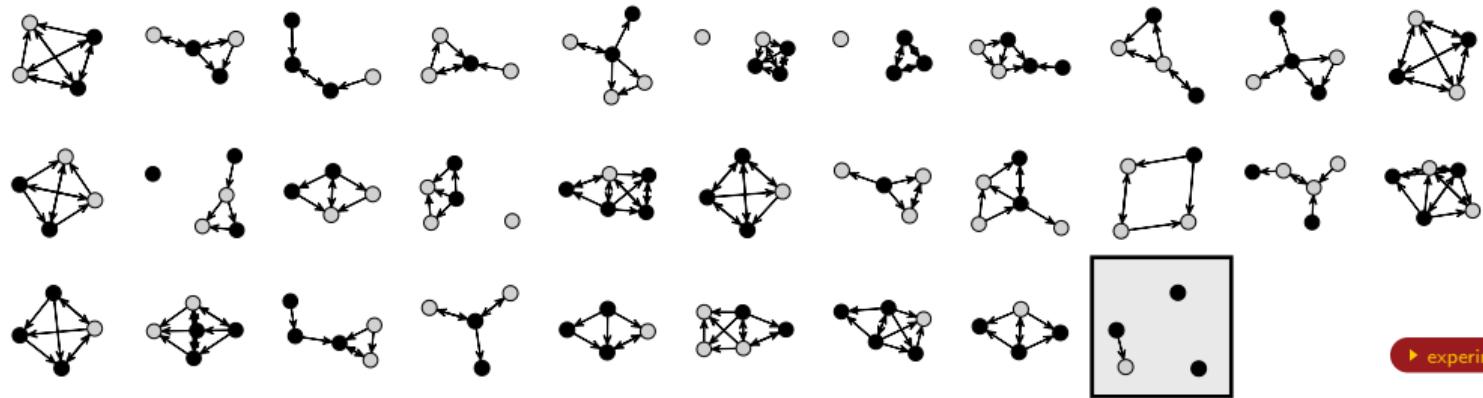


▶ experiment



▶ experiment

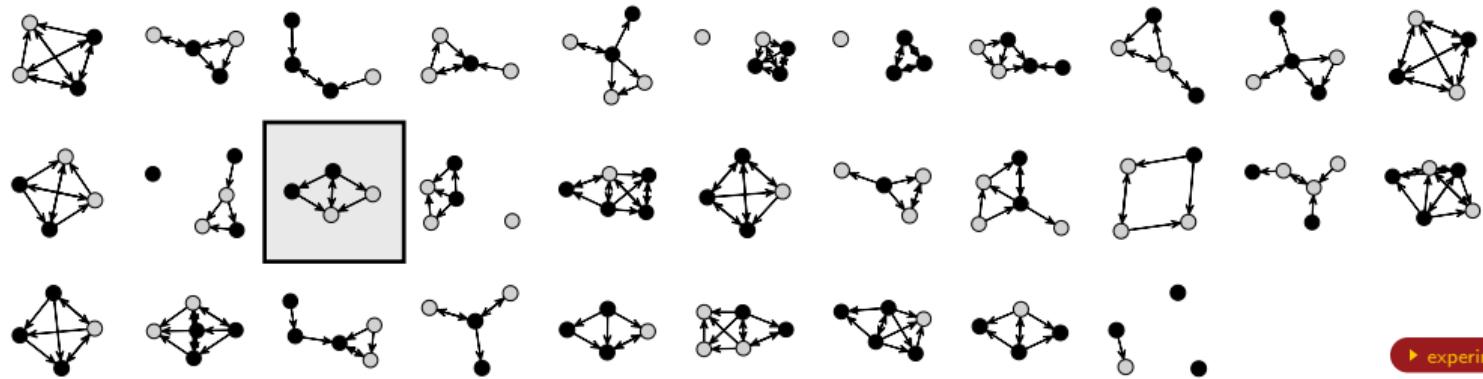
Key findings



▶ experiment

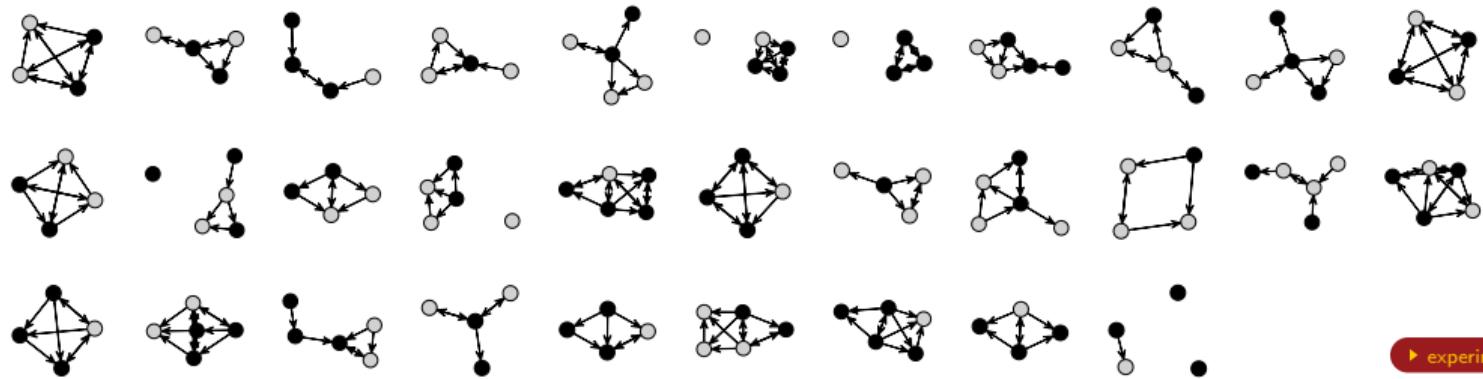
Key findings

- ▶ Low density.



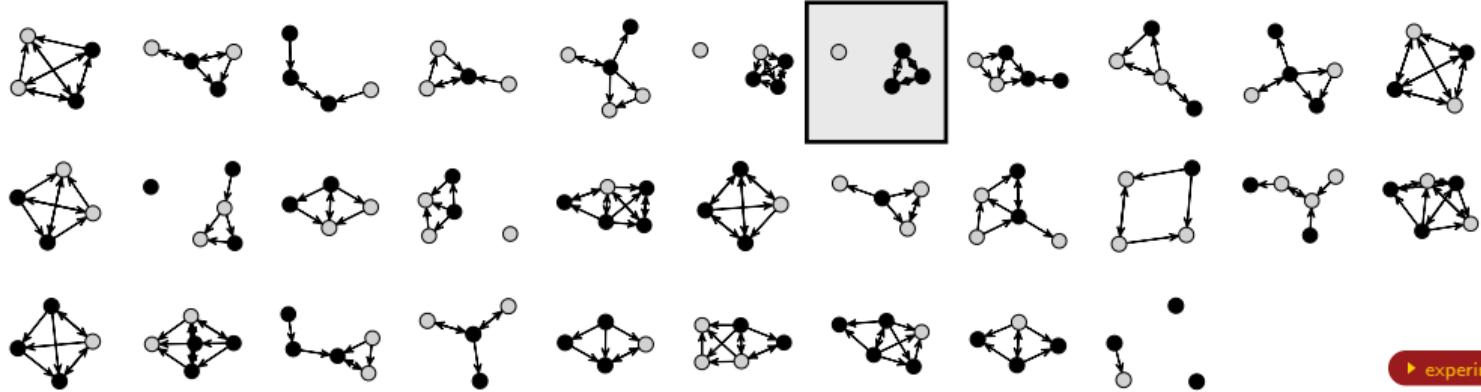
Key findings

- ▶ Low density.
- ▶ High balance (transitive triads).



Key findings

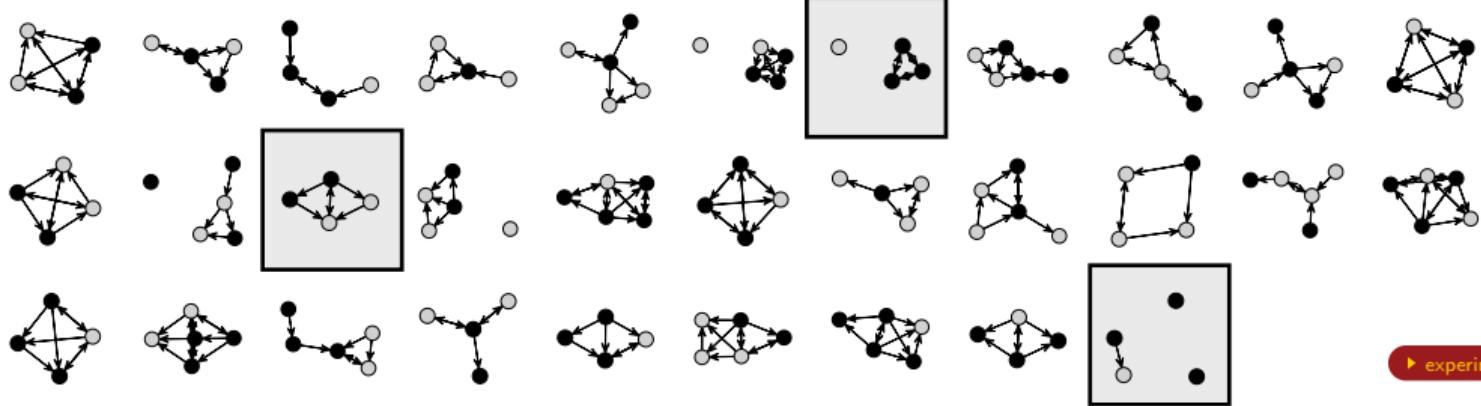
- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.



▶ experiment

Key findings

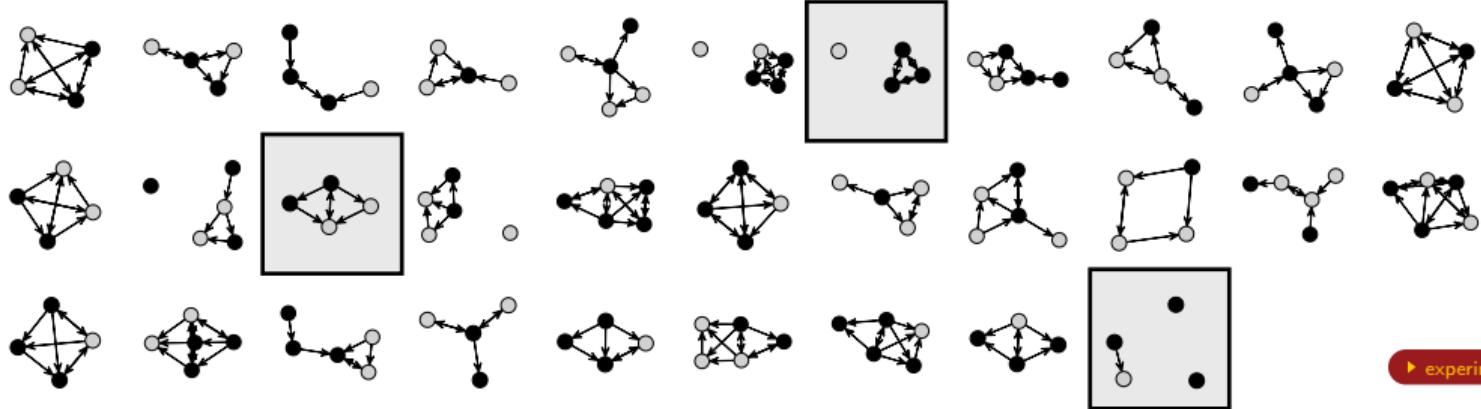
- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.
- ▶ Females are more likely to ask for advice.



Key findings

- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.
- ▶ Females are more likely to ask for advice.

What's different (from regular ERGMs) ?



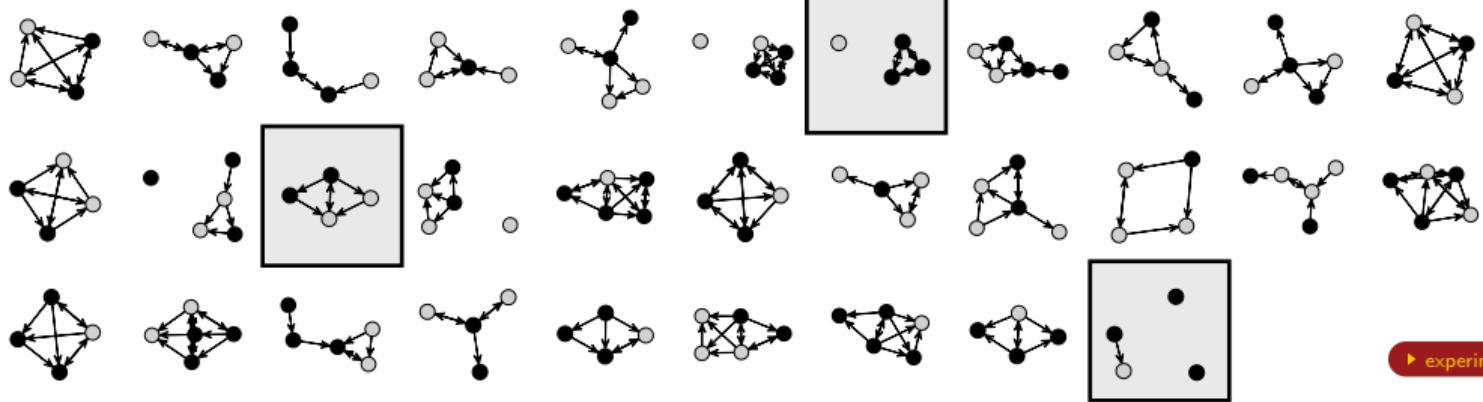
▶ experiment

Key findings

- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.
- ▶ Females are more likely to ask for advice.

What's different (from regular ERGMs)?

- ▶ Interaction effects: edges $\times \mathbf{1}$ ($n = 5$).

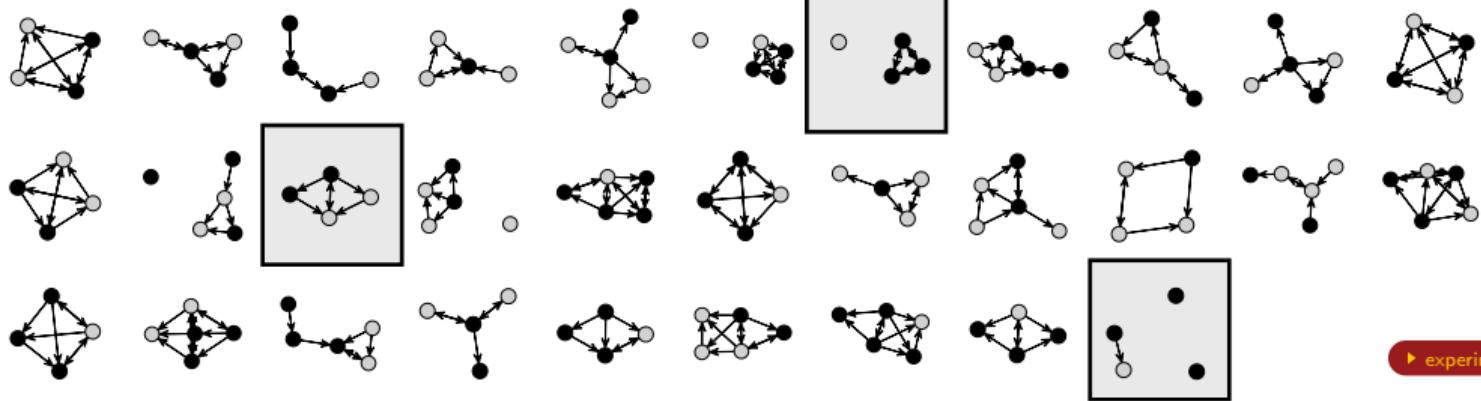


Key findings

- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.
- ▶ Females are more likely to ask for advice.

What's different (from regular ERGMs) ?

- ▶ Interaction effects: edges \times 1 ($n = 5$).
- ▶ Constrained support: edge > 4.

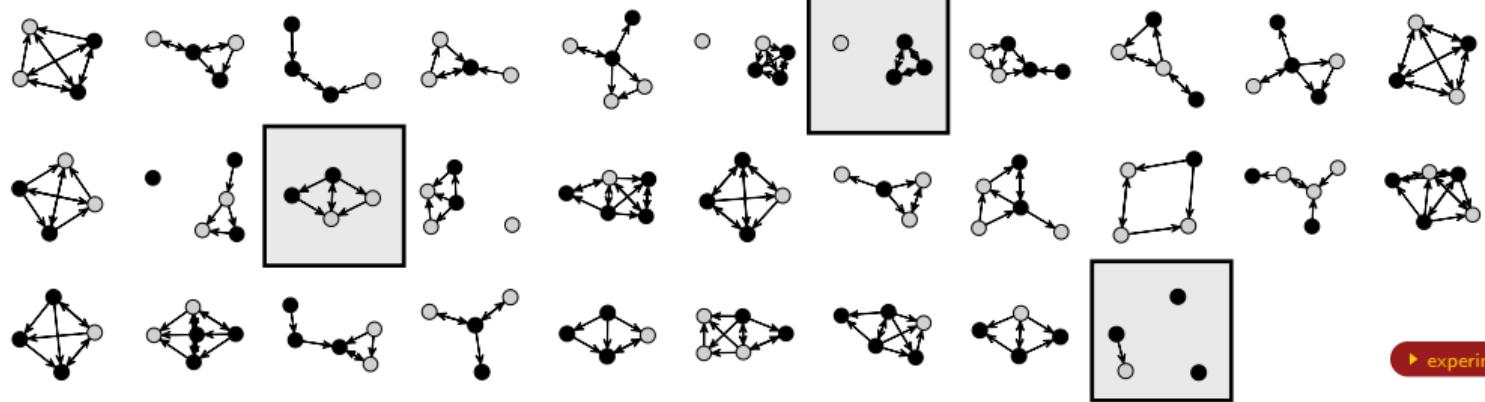


Key findings

- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.
- ▶ Females are more likely to ask for advice.

What's different (from regular ERGMs)?

- ▶ Interaction effects: edges $\times \mathbf{1}$ ($n = 5$).
- ▶ Constrained support: edge > 4 .
- ▶ Transformed variables: $(\text{Homophily})^{1/2}$.

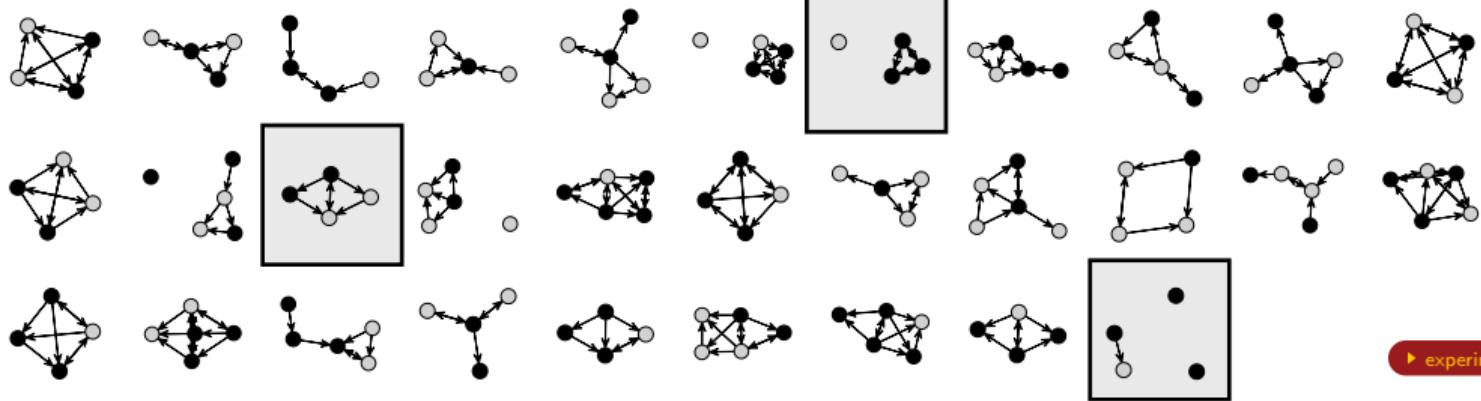


Key findings

- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.
- ▶ Females are more likely to ask for advice.

What's different (from regular ERGMs)?

- ▶ Interaction effects: edges $\times \mathbf{1}$ ($n = 5$).
- ▶ Constrained support: edge > 4 .
- ▶ Transformed variables: $(\text{Homophily})^{1/2}$.
- ▶ Bootstrapping: 1,000 replicates in less than 1.5 minutes...



Key findings

- ▶ Low density.
- ▶ High balance (transitive triads).
- ▶ No evidence of gender homophily.
- ▶ Females are more likely to ask for advice.

What's different (from regular ERGMs)?

- ▶ Interaction effects: edges $\times \mathbf{1}$ ($n = 5$).
- ▶ Constrained support: edge > 4 .
- ▶ Transformed variables: $(\text{Homophily})^{1/2}$.
- ▶ Bootstrapping: 1,000 replicates in less than 1.5 minutes...
... if you are lucky, using “regular” ERGMs would take you about 5 hours.

▶ details

▶ gof

▶ data

Computational Complexity: ergmito

Premise

Computational Complexity: ergmito

Premise

- ▶ Full enumeration \neq Large support

Computational Complexity: ergmito

Premise

- ▶ Full enumeration \neq Large support

Extra features

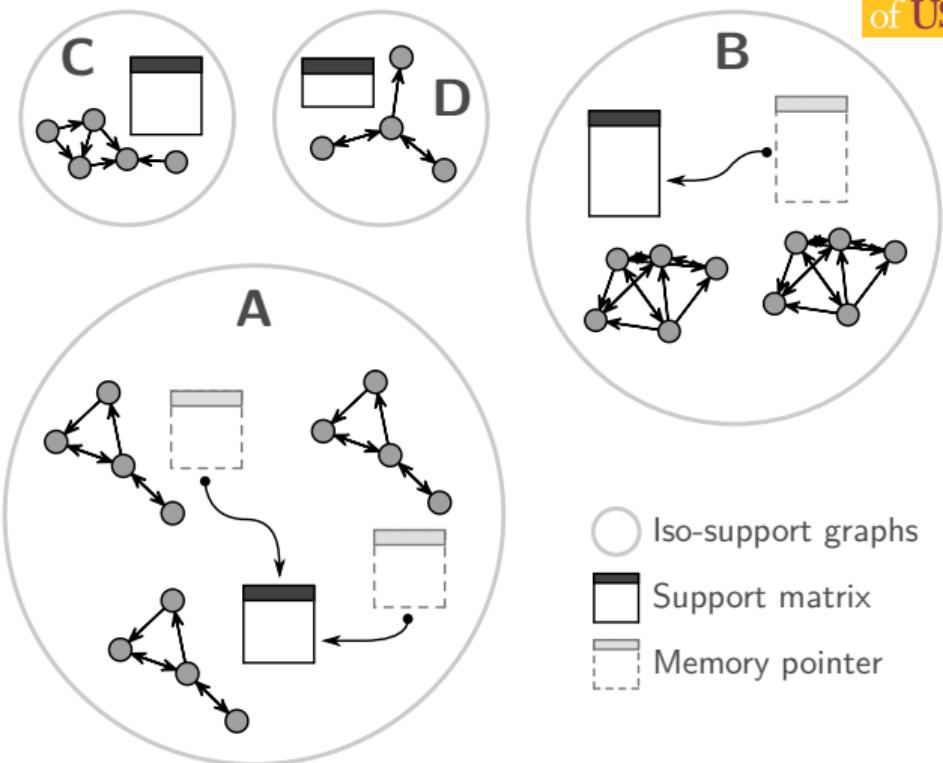
Computational Complexity: ergmito

Premise

- ▶ Full enumeration \neq Large support

Extra features

- ▶ *Iso-support* structures (model)
recycled



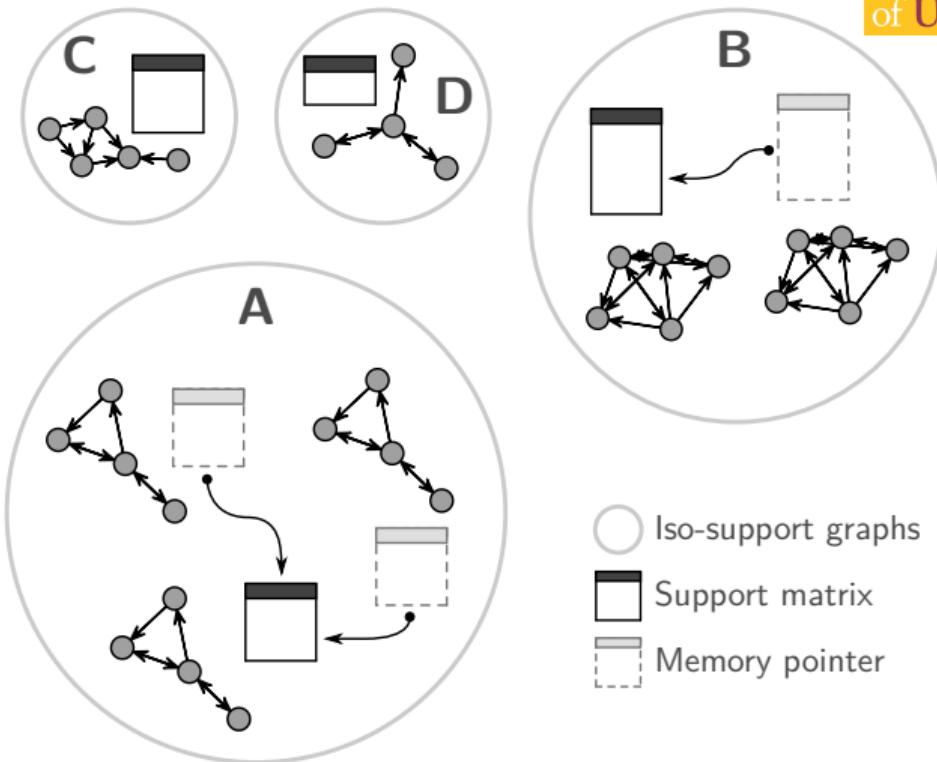
Computational Complexity: ergmito

Premise

- ▶ Full enumeration \neq Large support

Extra features

- ▶ *Iso-support* structures (model) recycled
- ▶ Normalizing constant updated as needed (e.g. BFGS)



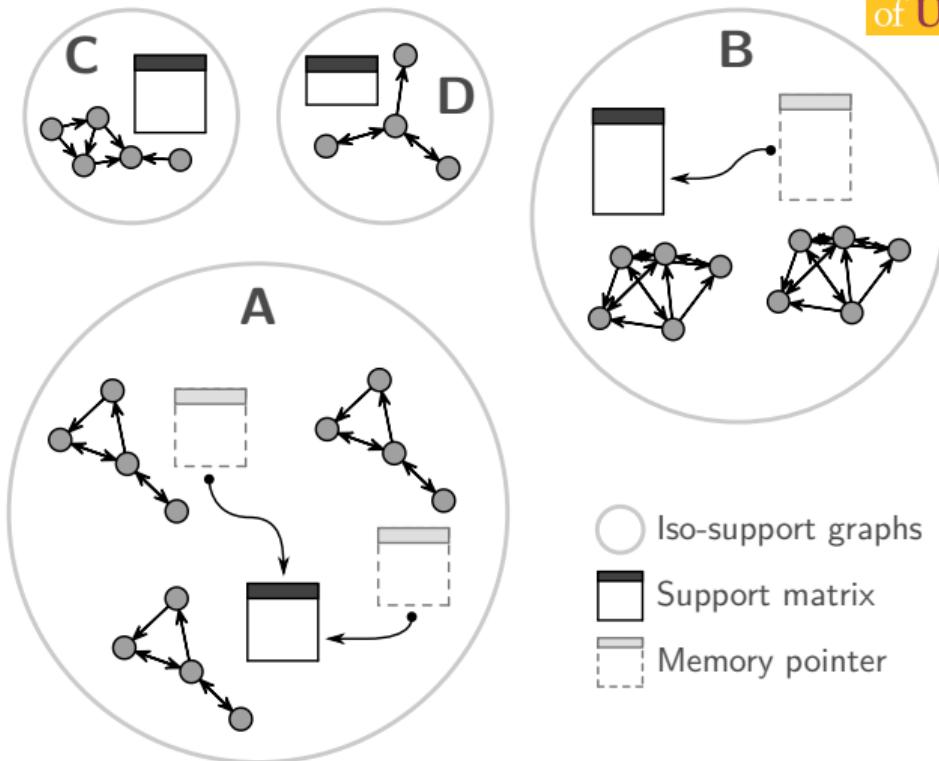
Computational Complexity: ergmito

Premise

- ▶ Full enumeration \neq Large support

Extra features

- ▶ *Iso-support* structures (model) recycled
- ▶ Normalizing constant updated as needed (e.g. BFGS)
- ▶ Core algorithm written in C++



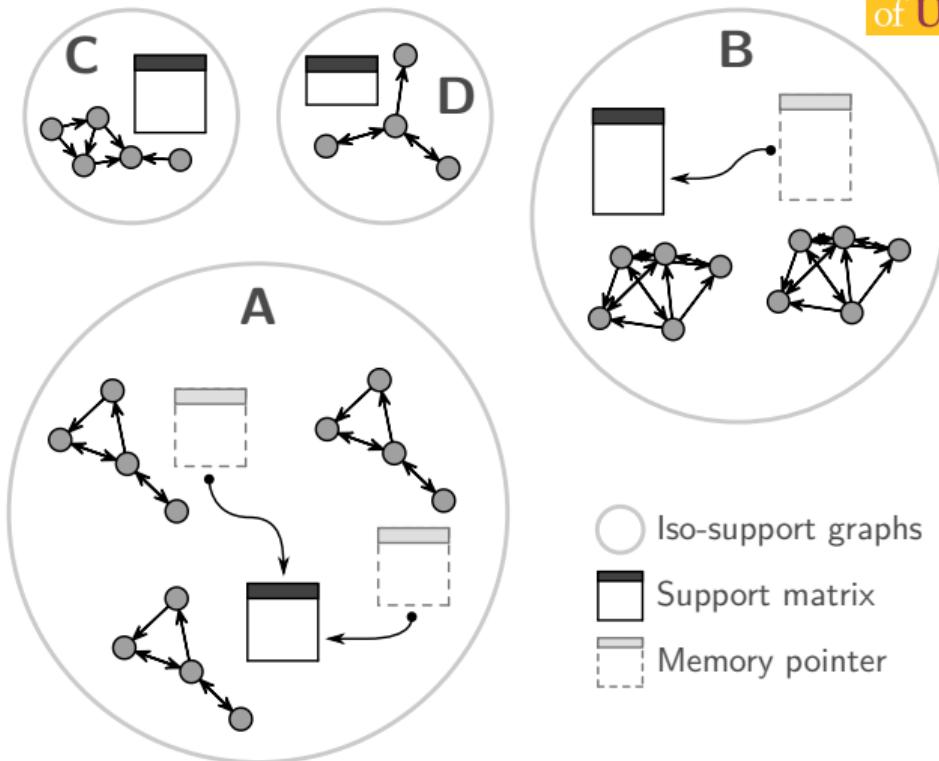
Computational Complexity: ergmito

Premise

- ▶ Full enumeration \neq Large support

Extra features

- ▶ *Iso-support* structures (model) recycled
- ▶ Normalizing constant updated as needed (e.g. BFGS)
- ▶ Core algorithm written in C++
- ▶ Parallel computing using **OpenMP**



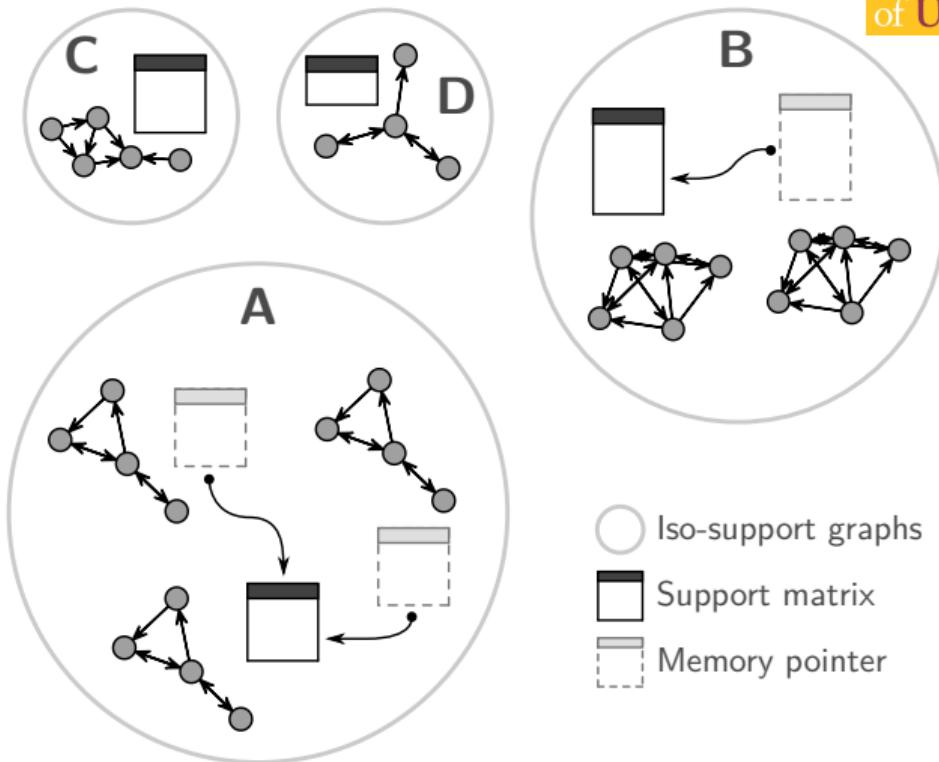
Computational Complexity: ergmito

Premise

- ▶ Full enumeration \neq Large support

Extra features

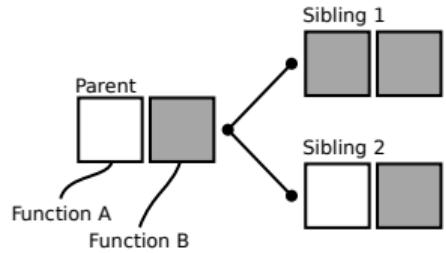
- ▶ *Iso-support* structures (model) recycled
- ▶ Normalizing constant updated as needed (e.g. BFGS)
- ▶ Core algorithm written in C++
- ▶ Parallel computing using **OpenMP**



Simple ERGMito with 80,000 nodes $\sim 4s$

Part 3: Current Research

Phylogenetics Modeling Strategies

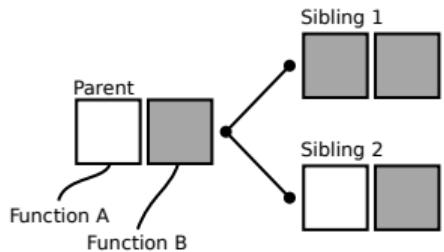


Has the function

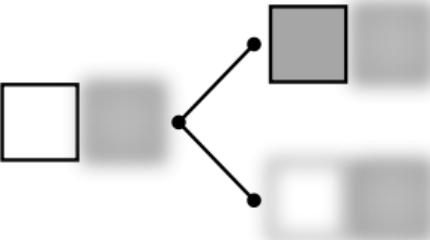


Doesn't have the function

Phylogenetics Modeling Strategies

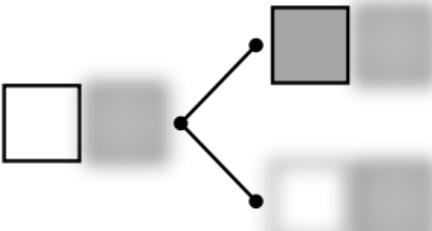
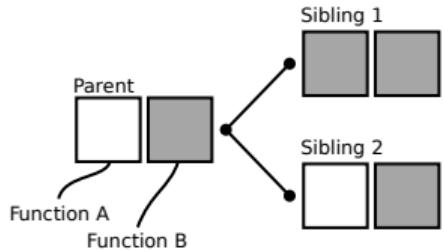


- [White Square] Has the function
- [Gray Square] Doesn't have the function

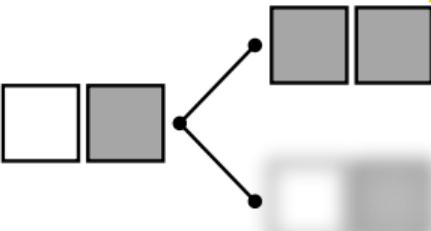


(a) Sibling and Function
Conditional Independence

Phylogenetics Modeling Strategies

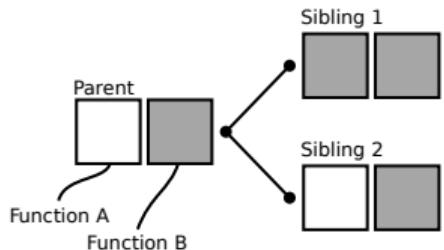


(a) Sibling and Function
Conditional Independence



(b) Sibling Conditional
Independence

- [White Square] Has the function
- [Gray Square] Doesn't have the function

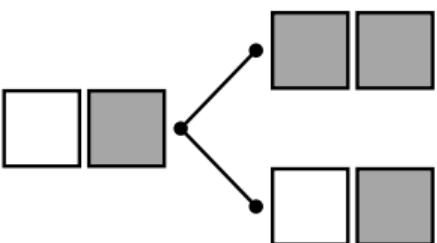


Has the function
 Doesn't have the function



(a) Sibling and Function
Conditional Independence

(b) Sibling Conditional
Independence



(c) No conditional
independence

If we wanted to build a model with 3 functions, we would need to estimate...

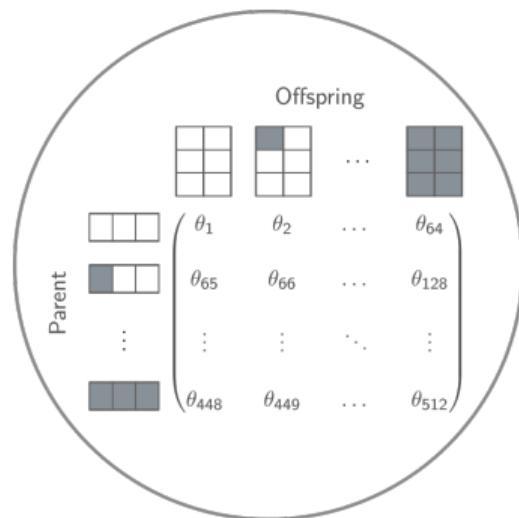
Full Markov Transition Matrix

Sufficient statistics

If we wanted to build a model with 3 functions, we would need to estimate...

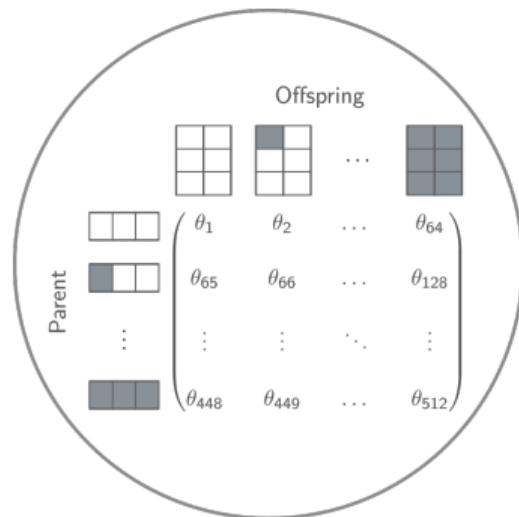
Full Markov Transition Matrix

Sufficient statistics

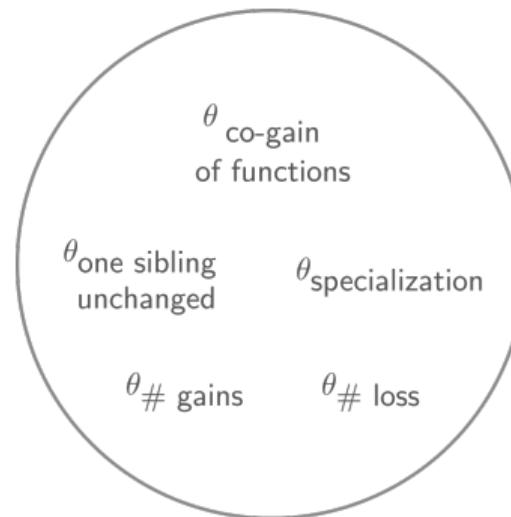


If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix

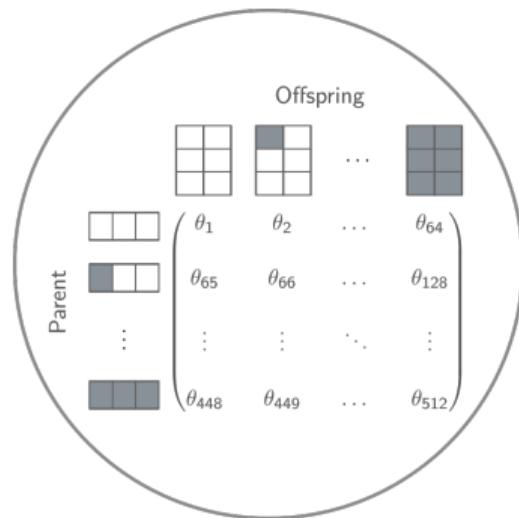


Sufficient statistics



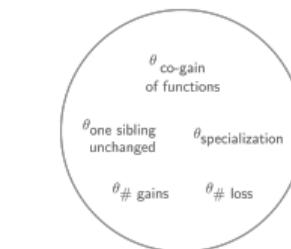
If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix

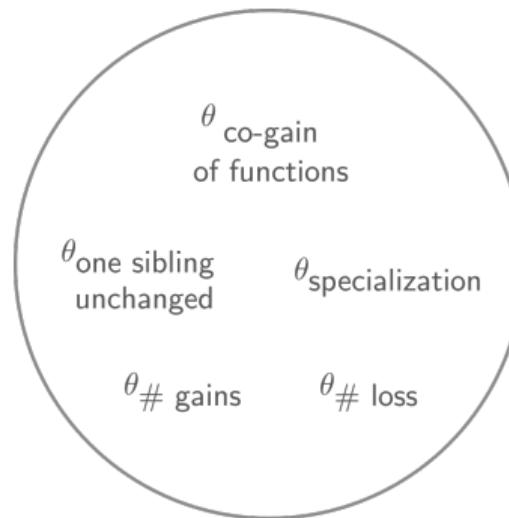


512 parameters

Sufficient statistics



Easier to fit
Easier to interpret



5 parameters

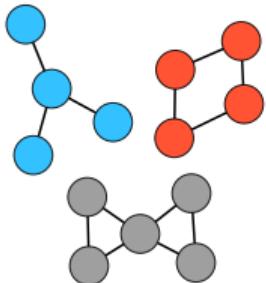
▶ numeric example

Part 4: Future Research

Step 1:

Fit the ERGMito

$$\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$$

**Step 2:**

Calculate $t_0 =$

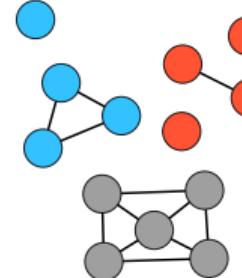
$$t \left(\begin{bmatrix} \text{blue motif} \\ \text{red motif} \\ \text{grey motif} \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \right)$$

Throughout the simulations
the only part that changes is
the networks, not Y

Step 3:

For $b \in 1, \dots, B$ do

$$\mathbf{G}^b = \{\mathbf{g}_1^b, \mathbf{g}_2^b, \mathbf{g}_3^b\}$$



Fit the ERGMito,

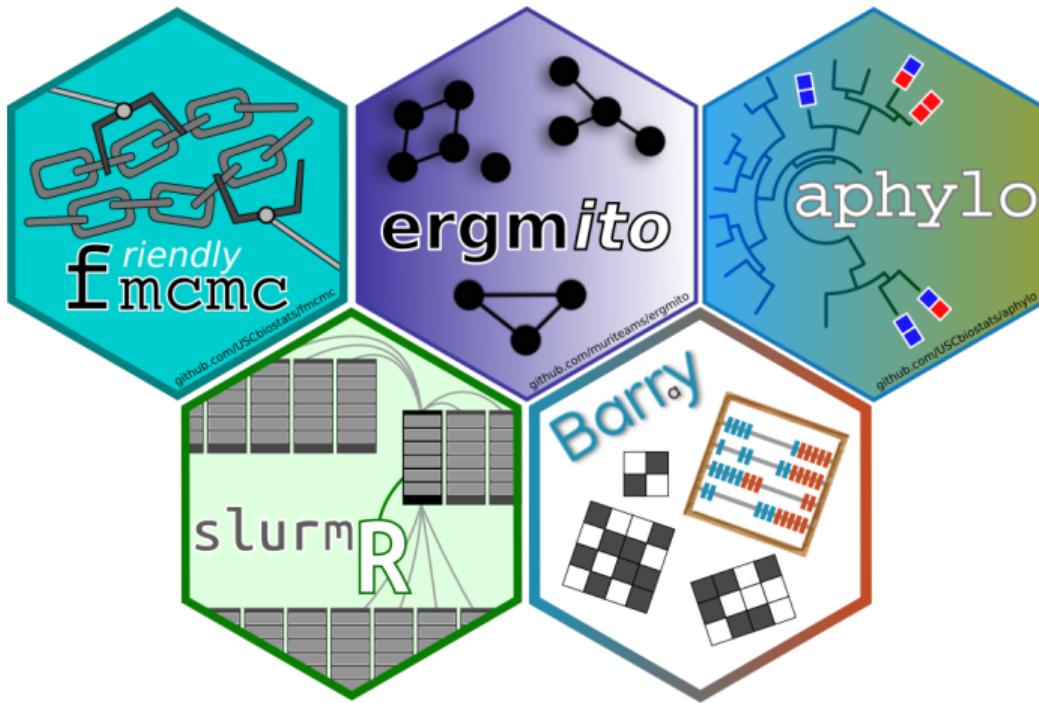
This will give us $\mathcal{D}(\hat{\theta}, X_j)$

3.1) For $j \in \{1, 2, 3\}$ draw a
new network from \mathcal{D}

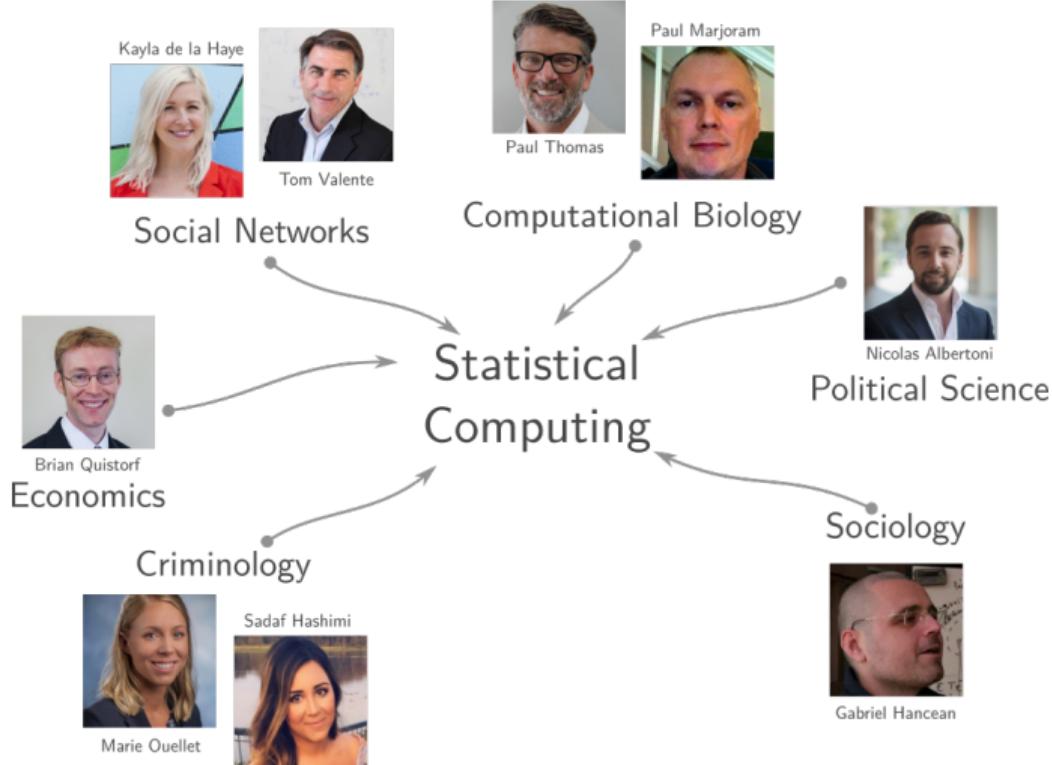
3.2) Use the new sample to
calculate $t_b = t(\mathbf{G}^b, Y)$

Figure 2 Semi-parametric test of association between graph structure, $t(\cdot)$ and a group level outcome y . Using ERGMitos, we can generate null distribution for complex hypothesis involving motifs beyond density and degree sequences.

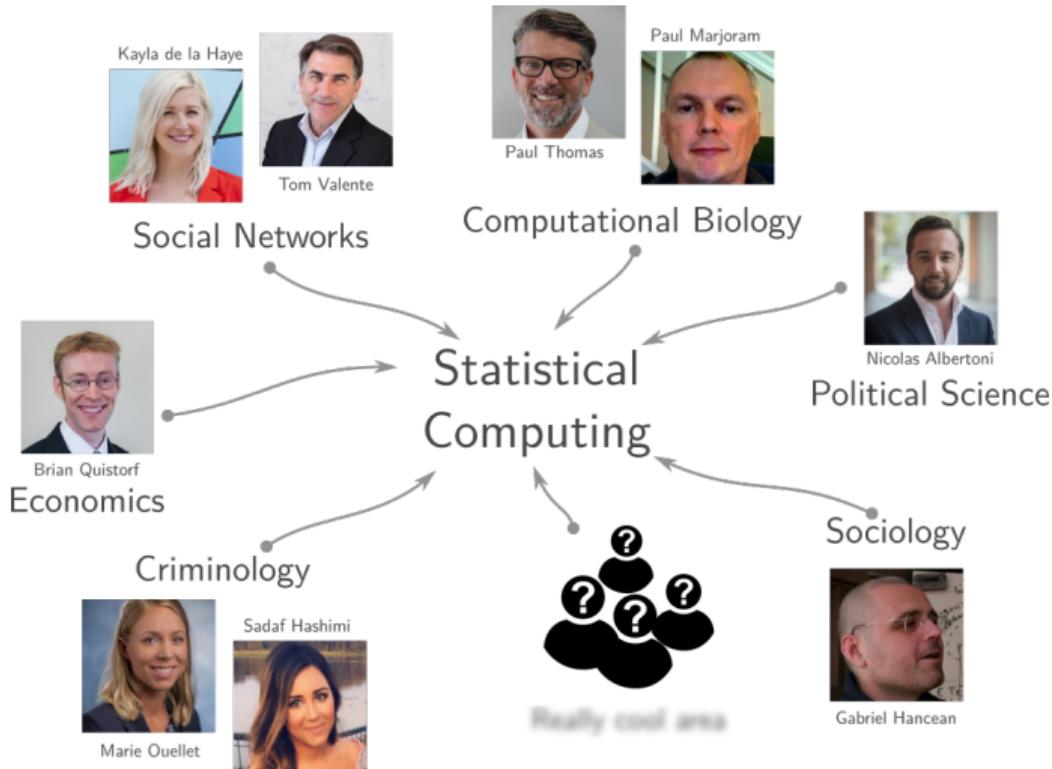
Continue developing scientific software...



Keep building bridges...



Keep building bridges...



...you could be next!

Statistical and Computational Methods for Complex Systems

George G Vega Yon

<https://ggyv.cl>

vegayon@usc.edu

Thank you!

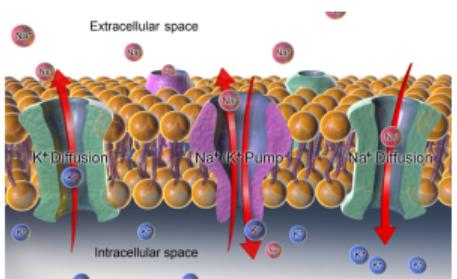
References |

-  Dodd, Diane M. B. (1989). "Reproductive Isolation as a Consequence of Adaptive Divergence in *Drosophila pseudoobscura*". In: Evolution 43.6, pp. 1308–1311. ISSN: 00143820, 15585646. URL: <http://www.jstor.org/stable/2409365>.
-  Handcock, Mark S. (2003). "Assessing Degeneracy in Statistical Models of Social Networks". In: Working Paper No. 39 76.39, pp. 33–50. ISSN: 1936900X. DOI: 10.1.1.81.5086. URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.81.5086>.

Gene functions can be classified in three types:

Molecular function

Active transport GO:0005215



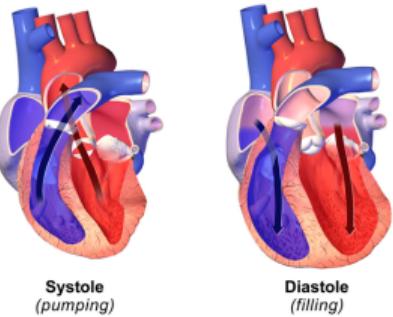
Cellular component

Mitochondria GO:0004016



Biological process

Heart contraction GO:0060047



◀ go back

The Gene Ontology Project

Example of GO term

Accession	GO:0060047
Name	heart contraction
Ontology	biological_process
Synonyms	heart beating, cardiac contraction, hemolymph circulation
Alternate IDs	None
Definition	The multicellular organismal process in which the heart decreases in volume in a characteristic way to propel blood through the body. Source: GOC:dph

Table 1 Heart Contraction Function. source: amigo.geneontology.org

You know what is interesting about this function?

◀ go back

These four species have a gene with that function... and two of these are part of the same evolutionary tree!



Felis catus pthr10037



Oryzias latipes pthr11521



Anolis carolinensis pthr11521



Equus caballus pthr24356

◀ go back

Example of Data + Predictions

Family: PTHR11258

Type: Molecular Function

Name: 2'-5'-oligoadenylate synthetase activity

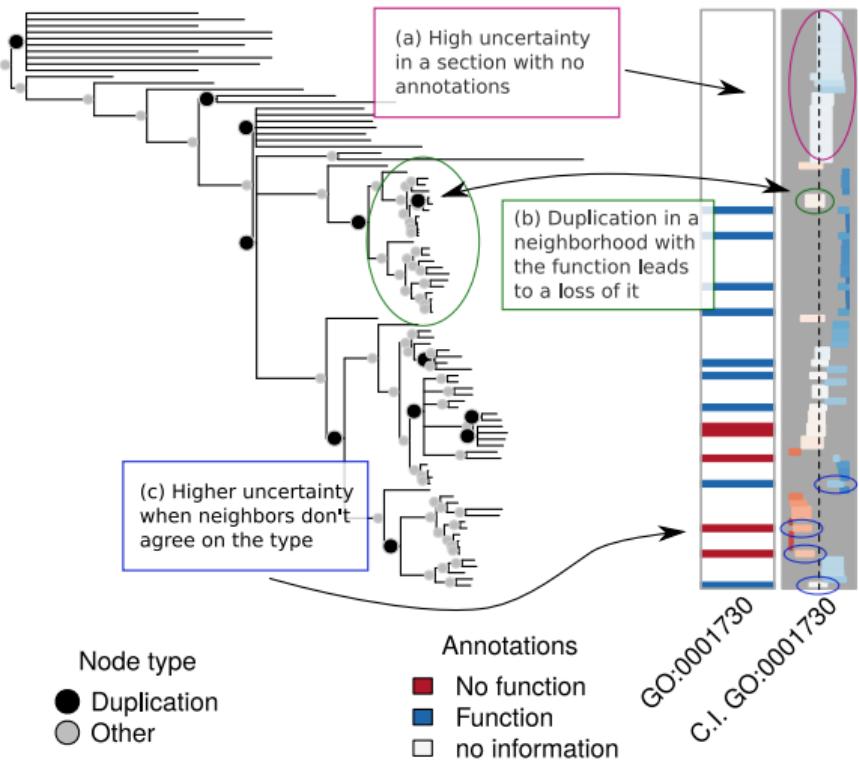
Desc: GO:0001730 involved in the process of cellular antiviral activity (wiki on [interferon](#)).

MAE: 0.34

AUC: 0.91

[see a bad one](#)

[◀ go back](#)

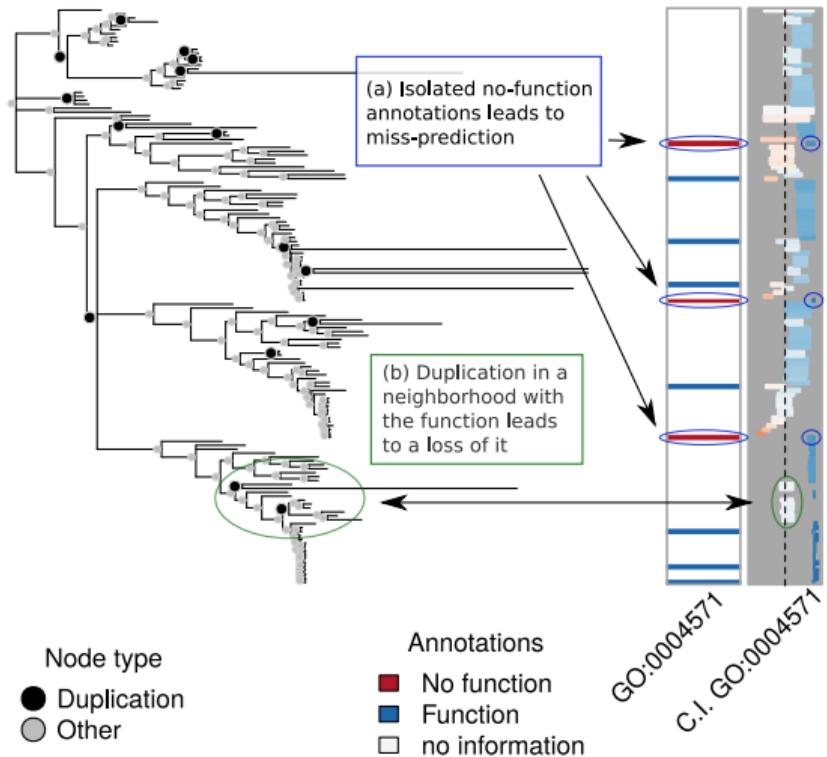


Example 2: Bad quality prediction

MAE: 0.52

AUC: 0.33

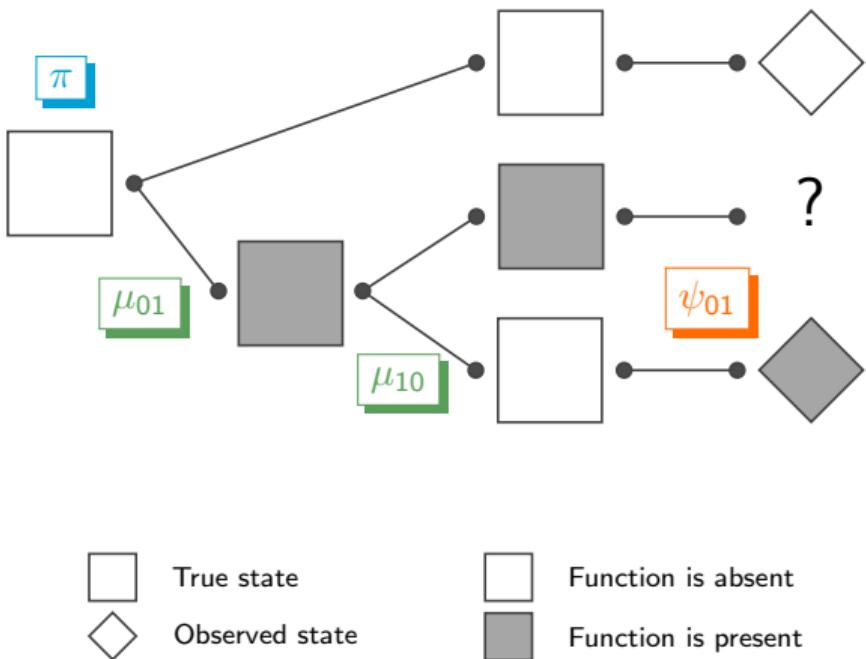
Type: Molecular Function

Name: mannosyl-oligosaccharide
1,2-alpha-mannosidase activityDesc: GO:0004571 involved in
synthesis of glycoproteins ([wiki](#)
and [examples](#)).[◀ go back](#)

		Pooled-data	One-at-a-time	
		Beta prior	Unif. prior	Beta Prior
Pooled-data				
Unif. prior	Beta prior	[-0.02,-0.01]	[-0.14,-0.10]	[-0.06,-0.03]
	Beta prior	-	[-0.12,-0.09]	[-0.04,-0.01]
One-at-a-time				
Unif. prior		-	-	[0.06, 0.09]

Table 2 Differences in Mean Absolute Error [MAE]. Each cell shows the 95% confidence interval for the difference in MAE resulting from two methods (row method minus column method). Cells are color coded blue when the method on that row has a significantly smaller MAE than the method on that column; Conversely, cells are colored red when the method in that column outperforms the method in that row. Overall, predictions calculated using the parameter estimates from *pooled-data* predictions outperform *one-at-a-time*.

An evolutionary model of gene functions



- ▶ Root has the function.
- ▶ Gain and loss (also depends on the type of event [► more](#)).
- ▶ Observed annotations may be incorrect.
- ▶ Only a fraction of the known genes have some form of annotation.

◀ go back

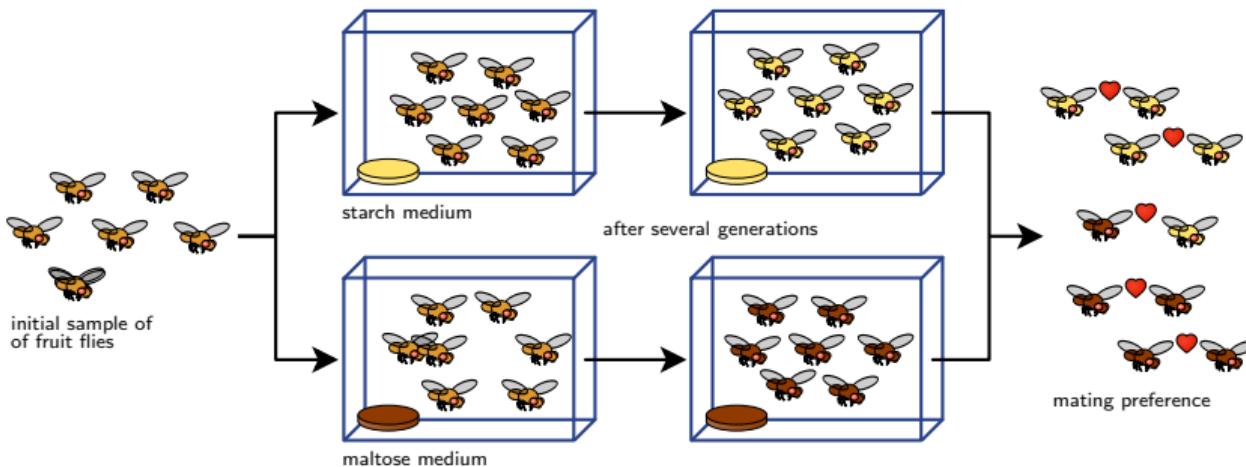


Figure 3 Dodd 1989: After one year of isolation, flies showed a significant level of assortativity in mating (wikimedia)

◀ go back

Duplication

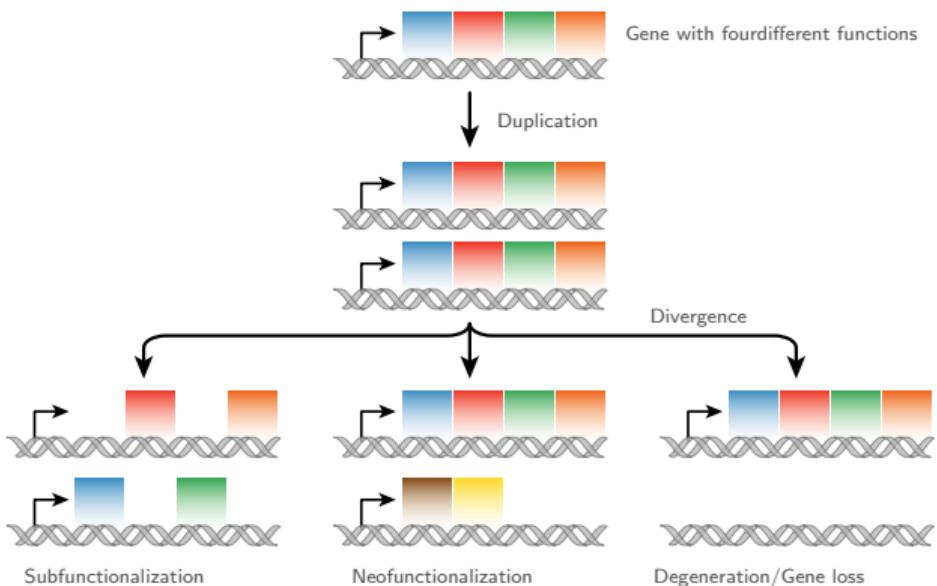


Figure 4 A key part of molecular innovation, gene duplication provides opportunity for new functions to emerge
(wikimedia)

◀ go back

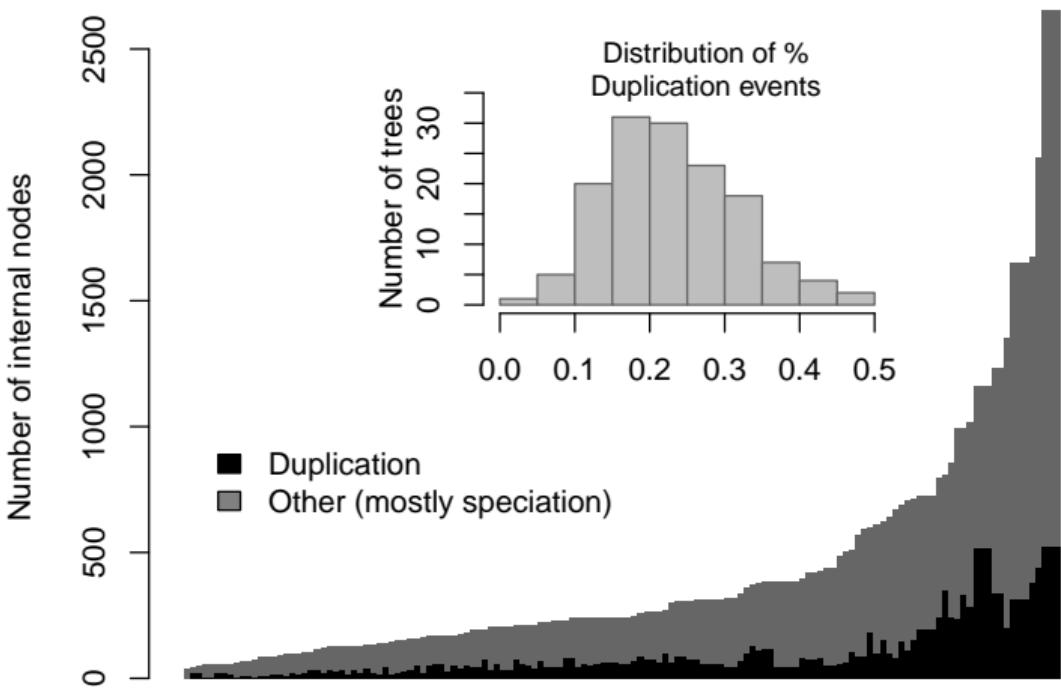
Data: Phylogenetic trees

Sample of annotations (first 10 in a single tree, Phosphoserine Phosphatase [PTHR10000])

Internal id	Branch Length	type	ancestor
AN0		S	LUCA
AN1	0.06	S	Archaea-Eukaryota
AN2	0.24	S	Eukaryota
AN3	0.44	S	Unikonts
AN4	0.42	S	Opisthokonts
AN6	0.68	D	
AN9	0.79	S	Amoebozoa
AN10	0.18	D	
AN15	0.57	S	Dictyostelium
AN18	0.52	S	Alveolata-Stramenopiles

◀ go back

Data: Node type (events)



◀ go back

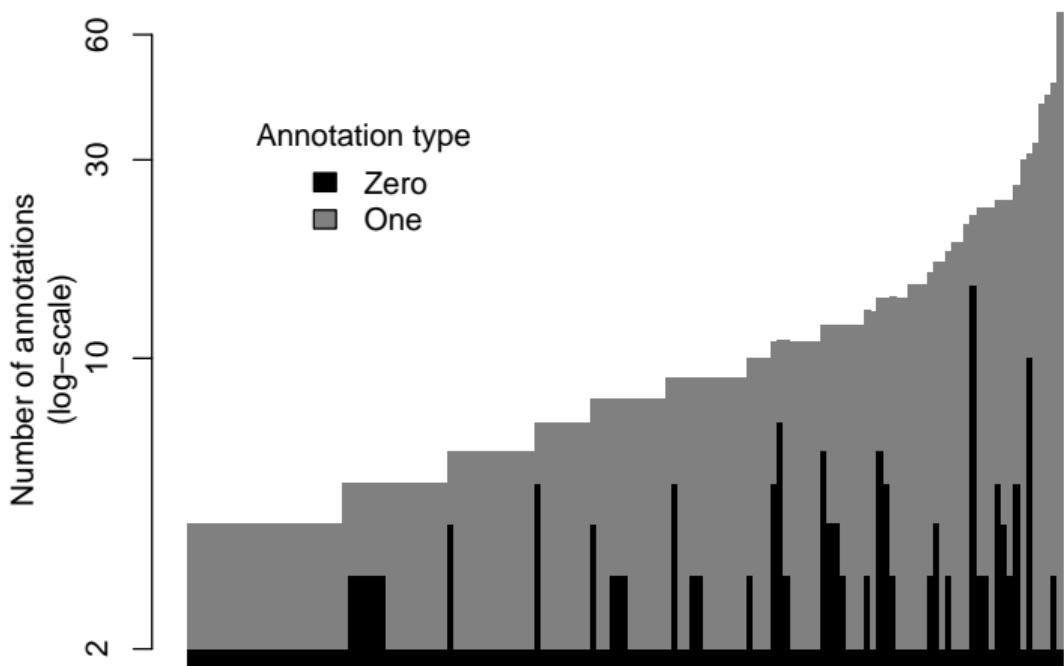
Data: Annotations (example)

This is the first 10 of ~ 400,000 experimental annotations used:

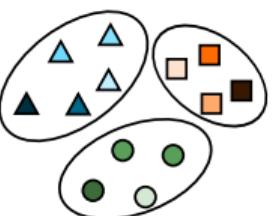
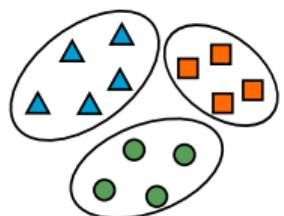
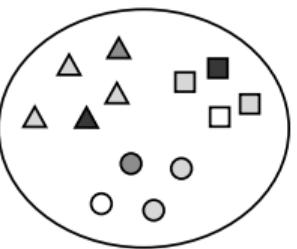
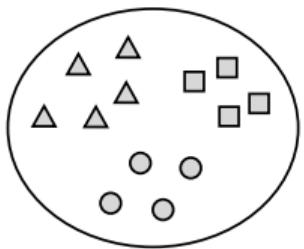
	Family	Id	GO term	Qualifier
1	PTHR12345	HUMAN HGNC=15756 UniProtKB=Q9H190	GO:0005546	
2	PTHR11361	HUMAN HGNC=7325 UniProtKB=P43246	GO:0016887	CONTRIBUTES_TO
3	PTHR10782	MOUSE MGI=MGI=3040693 UniProtKB=Q6P1E1	GO:0045582	
4	PTHR23086	ARATH TAIR=AT3G09920 UniProtKB=Q8L850	GO:0006520	
5	PTHR32061	RAT RGD=619819 UniProtKB=Q9EPI6	GO:0043197	
6	PTHR46870	ARATH TAIR=AT3G46870 UniProtKB=Q9STF9	GO:1990825	
7	PTHR15204	MOUSE MGI=MGI=1919439 UniProtKB=Q9Z1R2	GO:0045861	
8	PTHR22928	DROME FlyBase=FBgn0050085 UniProtKB=Q9XZ34	GO:0030174	
9	PTHR35972	HUMAN HGNC=34401 UniProtKB=A2RU48	GO:0005515	
10	PTHR10133	DROME FlyBase=FBgn0002905 UniProtKB=O18475	GO:0097681	

◀ go back

Data: Experimental Annotations



Phylogenetics Modeling: Pooling data



- (a) Featured in the first version of the model.
- (b) “Full glory” Hierarchical Bayes (1,001 parameters for the 141 functions).
- (c) Distilled version (a), improves accuracy.
- (d) Model estimated for Molecular Function (using Empirical Bayes) without significant improvements.

All methods are now available in the `aphylo` package: `aphylo_mle`, `aphylo_mcmc`, and `aphylo_hier`.

◀ go back

Overview of Prediction Results

	Pooled	Type of Annotation		
		Molecular Function	Biological Process	Cellular Comp.
Mislabeling				
ψ_{01}	0.23	0.18	0.09	
ψ_{10}	0.01	0.01	0.01	
Duplication Events				
μ_{d01}	0.97	0.97	0.10	
μ_{d10}	0.52	0.51	0.03	
Speciation Events				
μ_{s01}	0.05	0.05	0.05	
μ_{s10}	0.01	0.01	0.02	
Root node				
π	0.79	0.71	0.88	
Trees	141	74	45	22
Accuracy under the by-aspect model				
AUC	-	0.77	0.83	
MAE	-	0.34	0.26	
Accuracy under the pooled-data model				
AUC	-	0.77	0.75	
MAE	-	0.35	0.34	

Previously, joint estimates out-performed one-at-a-time

- ▶ **Molecular Function** No change.
- ▶ **Biological Process** Significantly better.
- ▶ **Cellular Component** Does not converge.

Molecular Function \neq Biological Process ? Cellular Component

▶ data

▶ go back

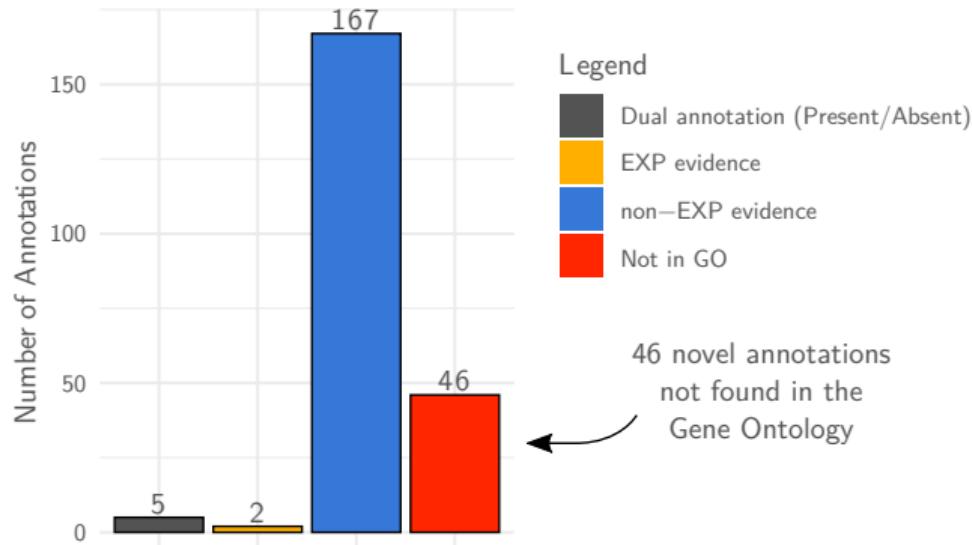


Figure 5 Distribution of predictions

◀ go back

Asymptotic Behavior of ERGMs

- ▶ In the case that $s_l = s(\mathbf{g}, x)$ is on the boundary: $s_l \rightarrow \pm\infty$
- ▶ Since the support space of $s(\mathbf{g}, x) \in \mathcal{S}$ is bounded, e.g. # edges $\in [0, n \times (n - 1)]$, we have:

$$\lim_{\theta_l \rightarrow \infty} l(\theta), \quad \lim_{\theta_l \rightarrow \infty} \nabla l(\theta), \quad \lim_{\theta_l \rightarrow \infty} \mathbf{H}(\theta)$$

log-likelihood, its gradient, and hessian are finite.

- ▶ The direct implication is that, while $s(\mathbf{g}, x)$ is on the boundary, the MLE for the other statistics exists.¹
- ▶ All equations ultimately involve realizations of $s(\mathbf{g}', x)$ that equal s_l , relevant in: Simulations, Bootstrapping, etc.

◀ go back

¹Handcock 2003 briefly mentions this

Conditional Graph Distributions

- ▶ Long history in (soc.) network science.
- ▶ Common usage: Hypothesis test prevalence of a feature.

Is the observed count of XYZ within the expected in a Bernoulli graph?

Are statistics A, B, and C different from graphs with 5 triangles?

- ▶ Different names, same thing, e.g. CUG tests and rewiring algorithms.
- ▶ $\{\text{CUG, Rewiring}\} \subset \text{ERGM}$
- ▶ We can talk about *Conditional ERGMs*.

$$\mathbb{P}(s(\mathbf{G})_k = s_k \mid s(\mathbf{G})_l = s_l, \theta) = \frac{\exp\{\theta_{-l}^t s(\mathbf{g})_{-l}\}}{\sum_{\mathbf{g}' : s(\mathbf{g}')_l = s_l} \exp\{\theta_{-l}^t s(\mathbf{g}')_{-l}\}}$$

In this equation θ_l becomes a nuisance parameter.

◀ go back

Sufficient statistics have various forms

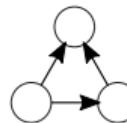
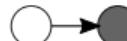
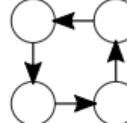
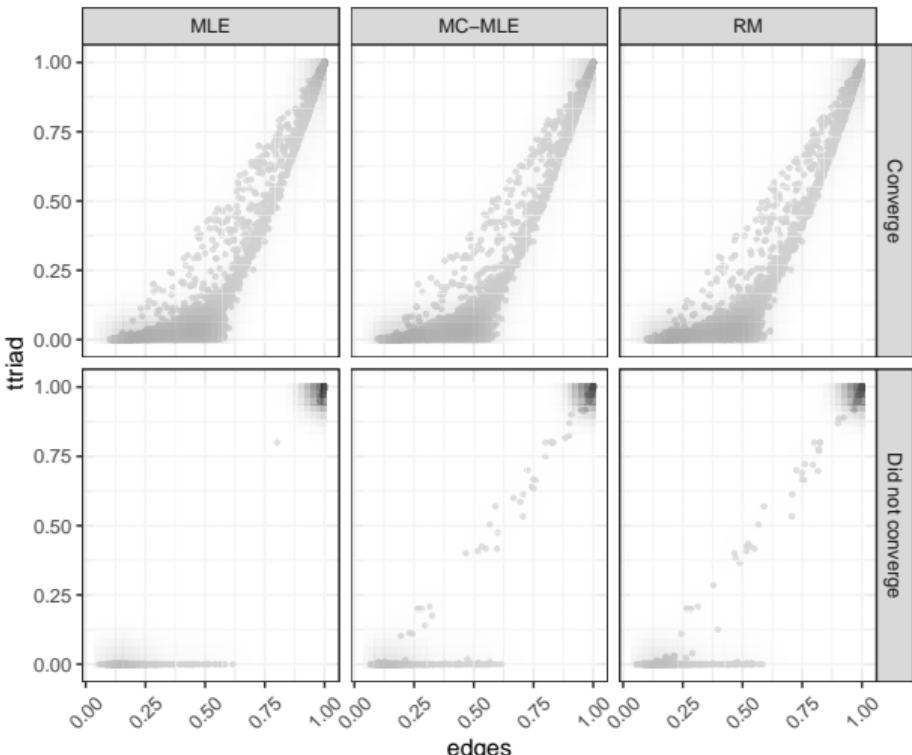
Representation	Description
	Mutual Ties (Reciprocity) $\sum_{i \neq j} y_{ij} y_{ji}$
	Transitive Triad (Balance) $\sum_{i \neq j \neq k} y_{ij} y_{jk} y_{ik}$
	Homophily $\sum_{i \neq j} y_{ij} \mathbf{1}(x_i = x_j)$
	Attribute-receiver effect $\sum_{i \neq j} y_{ij} x_j$
	Four Cycle $\sum_{i \neq j \neq k \neq l} y_{ij} y_{jk} y_{kl} y_{li}$

Figure 6 Besides the common edge count statistic (number of ties in a graph), ERGMs allow measuring other more complex structures that can be captured as sufficient statistics.

◀ go back

1. Higher convergence rate

◀ return



1. Higher convergence rate
2. **Smaller bias**

◀ return

	MLE	MC-MLE	RM
edges	[0.27, 0.36]	[1.23, 1.65]	[0.55, 1.54]
ttriads	[-0.05, -0.03]	[-0.22, -0.16]	[-0.15, 0.48]

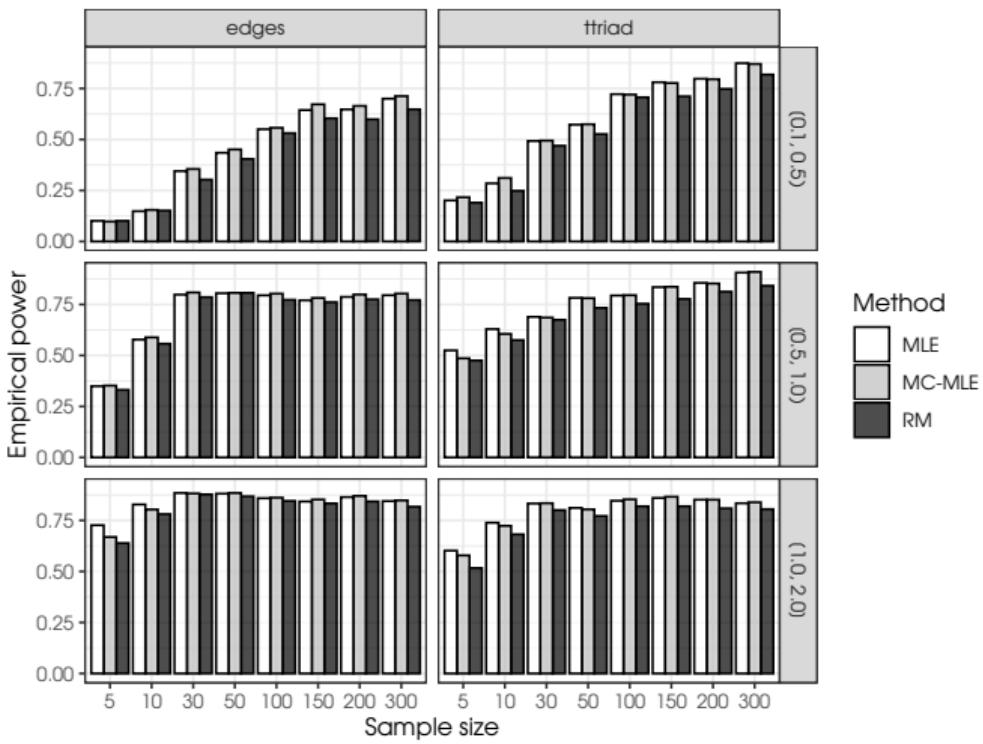
Table 3 Empirical bias. Each cell shows the 95% confidence interval of each methods' empirical bias.

▶ alt take

Simulation Study

1. Higher convergence rate
2. Smaller bias
3. **Higher power**

◀ return



Simulation Study

1. Higher convergence rate

2. Smaller bias

3. Higher power

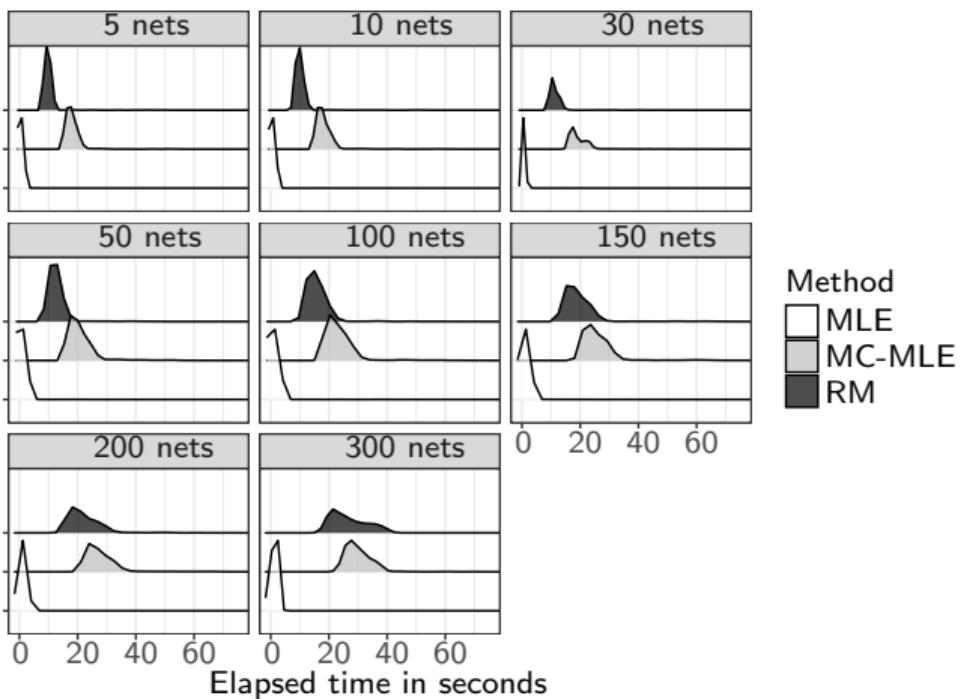
4. **Smaller type I error**[◀ return](#)

	Sample size	N. Sims.	P(Type I error)		χ^2 (vs MLE)		
			MLE	MC-MLE	RM	MC-MLE	RM
	5	4,325	0.066	0.086	0.086	11.36 ***	11.36 ***
	10	4,677	0.063	0.078	0.073	8.44 ***	3.73 *
	15	4,818	0.060	0.072	0.063	5.50 **	0.41
	20	4,889	0.054	0.065	0.061	5.30 **	2.05
	30	4,946	0.053	0.059	0.055	1.60	0.07
	50	4,987	0.053	0.055	0.047	0.16	1.67
	100	4,999	0.054	0.054	0.050	0.00	0.81

Simulation Study

1. Higher convergence rate
2. Smaller bias
3. Higher power
4. Smaller type I error
5. **Elapsed time**

◀ return



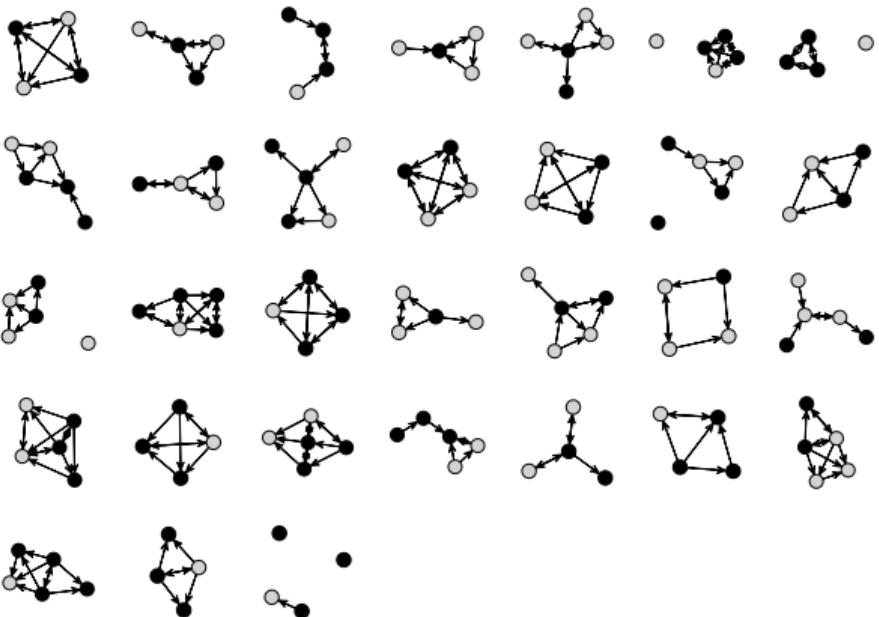
Featured Application: Small Teams

Sample

- ▶ 31 mixed-gender teams (4-5 members).
- ▶ University students.
- ▶ Don't know each other.
- ▶ At least one female/male per team.

Experiment

- ▶ Complete 1 hour of group tasks.
- ▶ Captured network data using name generator survey: *Who did you go to for advice, information or help to complete the group task?*



Is Gender Homophily a feature of these graphs?

◀ go back

	(1)	(2)	(3)	(4)	(5)	(4b)
edges	-0.52** (0.17)	-0.91*** (0.23)	-0.54** (0.18)	-0.72*** (0.19)	-0.48* (0.19)	-0.72*** (0.17)
ttriads	0.36*** (0.06)	0.46*** (0.06)	0.37*** (0.06)	0.36*** (0.06)	0.36*** (0.06)	0.36*** (0.05)
Homophily (gender)	-0.03 (0.20)	-0.01 (0.21)	-0.20 (0.46)	-0.12 (0.20)	-0.01 (0.20)	-0.12 (0.20)
edges × 1 ($n = 5$)	-0.53*** (0.12)	-0.47** (0.16)	-0.52*** (0.13)	-0.53*** (0.13)	-0.53*** (0.12)	-0.53*** (0.13)
(Homophily) $^{1/2}$			0.54 (1.32)			
Sender (female)				0.46* (0.18)		0.46* (0.18)
Receiver (female)					-0.08 (0.18)	
<i>Constraint (offset)</i>						
edge > 4		Yes				
AIC	639.26	569.93	641.08	634.68	641.07	634.68
BIC	655.99	586.66	661.99	655.59	661.98	655.59
Num. networks	31	28	31	31	31	31
Time (seconds)	2.26	2.32	2.28	5.10	5.19	83.97
N replicates					1000	

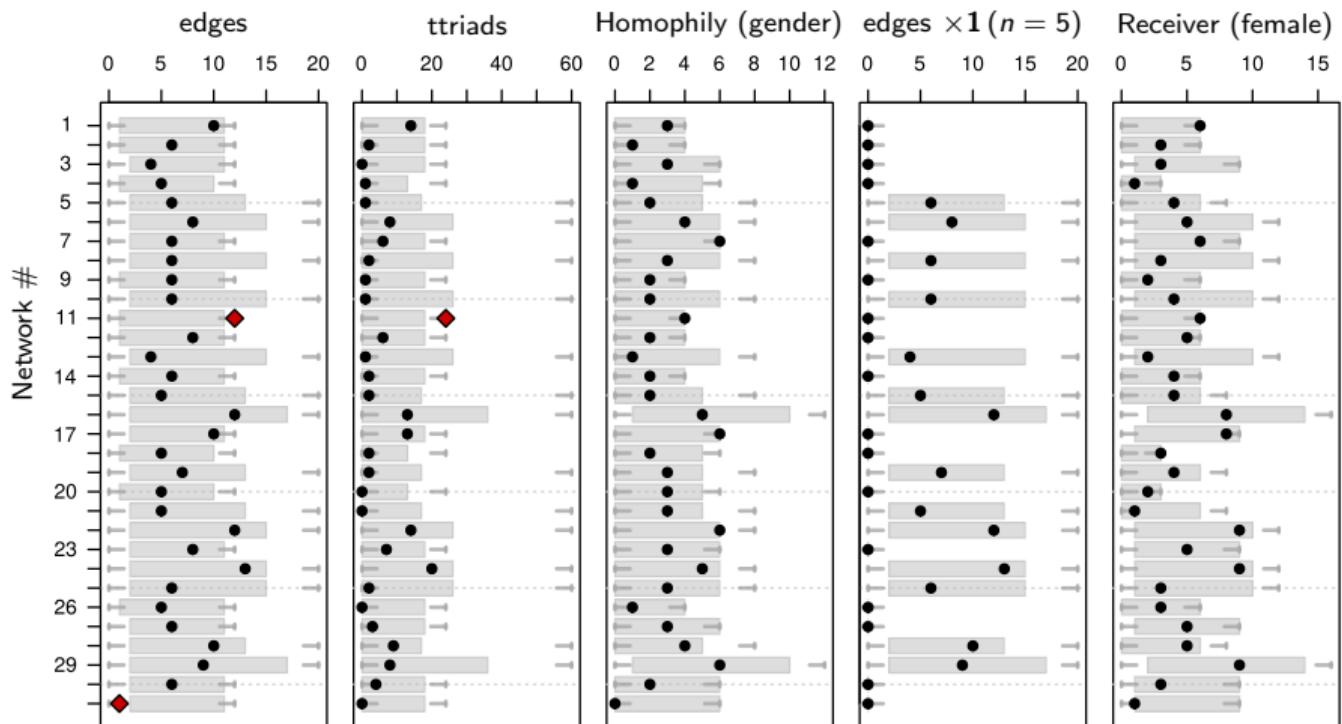
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

1. Interaction effects: seemingly included.
2. Transformed variables: also easy to add.
3. Using offset terms, we can constrain the support.
4. Each 1,000 bootstrap replicates took roughly 0.08 secs.
5. No support for gender homophily, but evidence of females sending more ties.

What about goodness-of-fit?

◀ go back

What About Goodness-of-fit?



◀ go back

(1)	(2)	(3)	(4)	(5)	(6)
Size (n)	edges	ttriads	edges \times $\mathbf{1} (n = 5)$	ttriads \times $\mathbf{1} (n = 5)$	edges \times $\log \{1/n\}$
4	10	14	0	0	-13.86
4	6	2	0	0	-8.32
4	4	0	0	0	-5.55
5	6	1	6	1	-9.66
5	8	8	8	8	-12.88
5	6	2	6	2	-9.66
... 25 more rows...					

Table 4 Example of observed sufficient statistics for the team advice networks. Pooled-data ERGMs have multiple observed sufficient statistics (also known as target statistics). Furthermore, as shown here, we can manipulate common statistics as *edges* (2) and *ttriads* (3) to include, e.g. interaction effects (4) and (5), or more complex transformations, e.g. (6).

Barry: your go-to *motif* accountant

- ▶ Sparse matrix represented using double hashmaps (fast row/column access).
- ▶ Template implementation for flexible weights and metadata.
- ▶ Fast counting using change statistics (Ch. 4).
- ▶ Calculation of support for sufficient stats.

[https://USCbiostats.github.io/
binaryarrays](https://USCbiostats.github.io/binaryarrays)

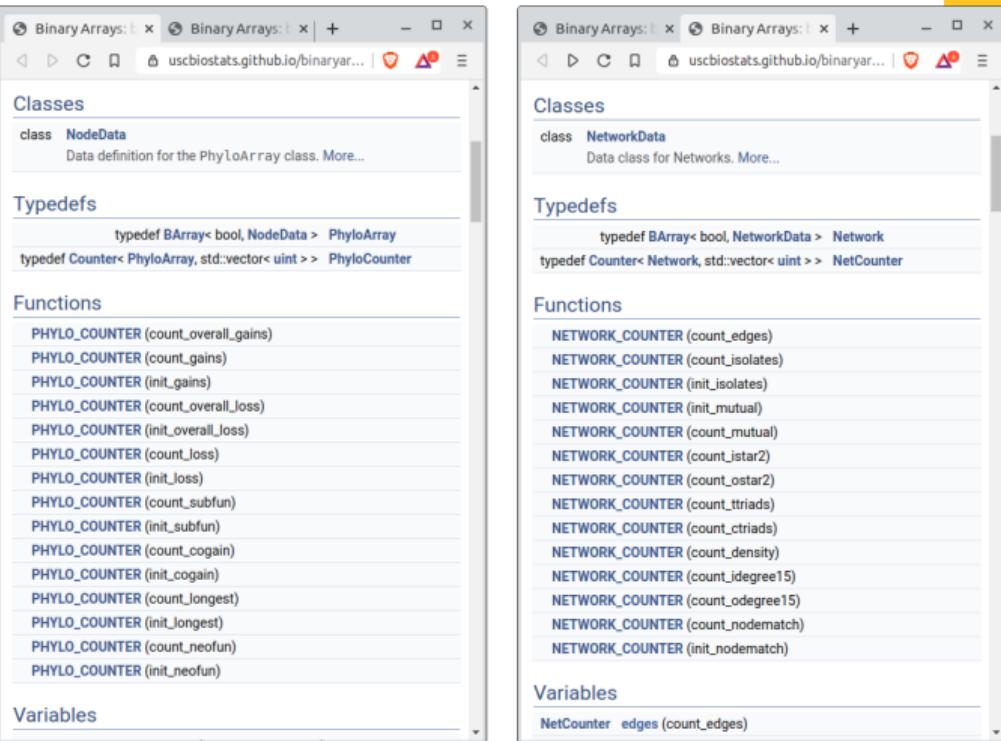


Figure 7 Screenshots from the project's website on GitHub.

What Drives Evolution

Imagine that we have 3 functions (rows) and that each node has 2 siblings (columns)

		Transitions to	
		Case 1	Case 2
Parent	A	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
	B	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$
	C	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$

Sufficient statistics

# Gains	1	1
Only one offspring changes (yes/no)	1	0
# Changes (gain+loss)	2	3
Subfunctionalizations (yes/no)	0	1

▶ return

What Drives Evolution: a game changer

In the model with 3 functions and 2 offspring per node:

- ▶ Full Markov transition matrix: $2^3 \times 2^6 = 512$
- ▶ Using sufficient statistics:

Pairwise co-evolution: 3 terms,

Pairwise Neofunctionalization: 3 terms,

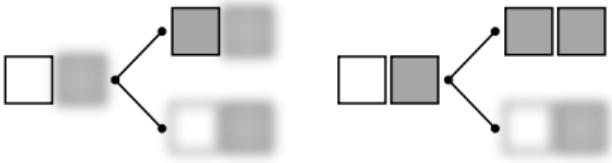
Pairwise Subfunctionalization: 3 terms,

Function specific gain: 3 terms,

Function specific loss: 3 terms,

Total: 15 parameters.

- ▶ Easier to fit and interpret.



(b) Sibling Conditional Independence

