Essays on Bioinformatics and Social Network Analysis

Statistical and Computational Methods for Complex Systems

George G Vega Yon

University of Southern California, Department of Preventive Medicine

January 28, 2020

Keck School of Medicine of USC

What motivates my research

Statistical and computational methods for bioinformatics and social network analysis

- We live in a non-*IID* world.

What motivates my research

Keck School of Medicine of USC

Statistical and computational methods for bioinformatics and social network analysis

- We live in a non-*IID* world.
- In some times, the cannot understand a process unless we look at it as a whole.

What motivates my research

Keck School of Medicine of USC

Statistical and computational methods for bioinformatics and social network analysis

- We live in a non-*IID* world.
- In some times, the cannot understand a process unless we look at it as a whole.
- There's a reason why we usually assume *IID*.

What motivates my research

Keck School of Medicine of USC

Statistical and computational methods for bioinformatics and social network analysis

- We live in a non-*IID* world.
- In some times, the cannot understand a process unless we look at it as a whole.
- There's a reason why we usually assume *IID*.
- *Modern* (as of today) computational tools help us coping with that.

Contents

Exponential Random Graph Models for Small Networks

Joint with: Andrew Slaughter and Kayla de la Haye

What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka ERGMs are:

What are Exponential Random Graph Models

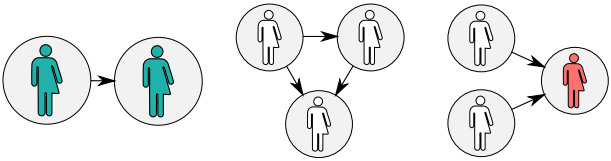
Exponential Family Random Graph Models, aka ERGMs are:

- Statistical models of (social) networks

What are Exponential Random Graph Models

Exponential Family Random Graph Models, aka ERGMs are:

- Statistical models of (social) networks
- In simple terms: statistical inference on what network patterns/structures/motifs govern social networks



ERGMs (cont'd)

A vector of model parameters

A vector of sufficient statistics

$$\Pr(\mathbf{Y} = \mathbf{y} \mid \theta, \mathbf{X}) = \frac{\exp\{\theta^{\mathbf{t}} s(\mathbf{y}, \mathbf{X})\}}{\sum_{\mathbf{y}' \in \mathcal{Y}} \exp\{\theta^{\mathbf{t}} s(\mathbf{y}', \mathbf{X})\}}, \quad \forall \mathbf{y} \in \mathcal{Y}$$

Observed data

The normalizing constant

All possible networks

ERGMs (cont'd)

A vector of model parameters

A vector of sufficient statistics

$$\Pr\left(\mathbf{Y} = \mathbf{y} \mid \theta, \mathbf{X}\right) = \frac{\exp\left\{\theta^{\mathbf{t}} s\left(\mathbf{y}, \mathbf{X}\right)\right\}}{\sum_{\mathbf{y}' \in \mathcal{Y}} \exp\left\{\theta^{\mathbf{t}} s\left(\mathbf{y}', \mathbf{X}\right)\right\}}, \quad \forall \mathbf{y} \in \mathcal{Y}$$

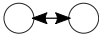
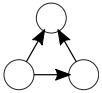
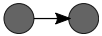
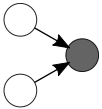
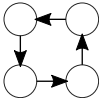
Observed data

The normalizing constant

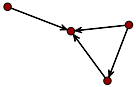
All possible networks

The normalizing constant has $2^{n(n-1)}$ terms!

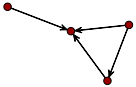
Sufficient statistics have various forms

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 $\sum_{i \neq j} y_{ij} \mathbf{1}(x_i = x_j)$
	Covariate Effect for Incoming Ties $\sum_{i \neq j} y_{ij} x_j$
	Four Cycle $\sum_{i \neq j \neq k \neq l} y_{ij} y_{jk} y_{kl} y_{li}$

In this network

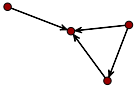


In this network



We see 4 **edges**, 1 **transitive triad** and **no mutual ties**.

In this network



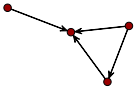
We see 4 **edges**, 1 **transitive triad** and **no mutual ties**.

The probability function of this model would be

$$\mathbb{P}(\mathbf{G} = \mathbf{g} \mid \theta) = \frac{\exp\{4\theta_{edges} + \theta_{ttriads} + 0\theta_{mutual}\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \exp\{\theta^{\mathsf{t}} s(\mathbf{g}')\}}$$

with $\theta = [\theta_{edges} \quad \theta_{ttriads} \quad \theta_{mutual}]^{\mathsf{t}}$

In this network



We see 4 **edges**, 1 **transitive triad** and **no mutual ties**.

The probability function of this model would be

$$\mathbb{P}(\mathbf{G} = \mathbf{g} \mid \theta) = \frac{\exp\{4\theta_{edges} + \theta_{ttriads} + 0\theta_{mutual}\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \exp\{\theta^{\mathbf{t}} s(\mathbf{g}')\}}$$

with $\theta = [\theta_{edges} \quad \theta_{ttriads} \quad \theta_{mutual}]^{\mathbf{t}}$

This model has **MLE parameter estimates** of -0.20 (low density), 0.28 (high chance of ttriads), and -Inf (low chance of mutuality) for the parameters `edges`, `ttriads`, and `mutual` respectively.

ERGMs: State of the Art

ERGMs: State of the Art

Medium-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic Approximation. details
- Maximum Pseudo Likelihood (MPLE)

ERGMs: State of the Art

Medium-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic Approximation. details
- Maximum Pseudo Likelihood (MPLE)

large-huge networks (up to millions of vertices)

- Parametric bootstrap
- Conditional joint estimation (like snowball sampling, a.k.a. divide and conquer)
- Equilibrium Expectation Algorithm (millions of vertices)

ERGMs: State of the Art

Medium-large (dozens to a couple of thousand vertices) networks

- Markov Chain Monte Carlo (MCMC) based approaches like MC-MLE or Robbins-Monro Stochastic Approximation. details
- Maximum Pseudo Likelihood (MPLE)

large-huge networks (up to millions of vertices)

- Parametric bootstrap
- Conditional joint estimation (like snowball sampling, a.k.a. divide and conquer)
- Equilibrium Expectation Algorithm (millions of vertices)

All of these methods are approximations!

Do we care about small networks?

We see small networks everywhere

Do we care about small networks?

We see small networks everywhere

- Families and friends

Do we care about small networks?

We see small networks everywhere

- Families and friends
- Small teams

Do we care about small networks?

We see small networks everywhere

- Families and friends
- Small teams
- Egocentric networks

Do we care about small networks?

We see small networks everywhere

- Families and friends
- Small teams
- Egocentric networks
- Online networks (sometimes)

Do we care about small networks?

We see small networks everywhere

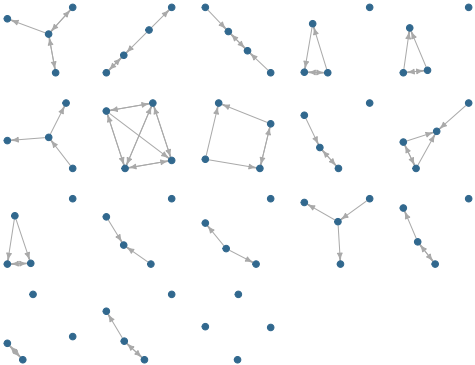
- Families and friends
- Small teams
- Egocentric networks
- Online networks (sometimes)
- etc.

Do we care about small networks?

We see small networks everywhere

- Families and friends
- Small teams
- Egocentric networks
- Online networks (sometimes)
- etc.

Small networks come in samples



ERGMs for Small Networks

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.

ERGMs for Small Networks

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- This allow us to directly compute **the normalizing constant**.

ERGMs for Small Networks

A vector of model parameters

A vector of sufficient statistics

$$\Pr(\mathbf{Y} = \mathbf{y} \mid \theta, \mathbf{X}) = \frac{\exp\{\theta^{\mathsf{t}} s(\mathbf{y}, \mathbf{X})\}}{\sum_{\mathbf{y}' \in \mathcal{Y}} \exp\{\theta^{\mathsf{t}} s(\mathbf{y}', \mathbf{X})\}}, \quad \forall \mathbf{y} \in \mathcal{Y}$$

Observed data

The normalizing constant

All possible networks

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- This allow us to directly compute **the normalizing constant**.

ERGMs for Small Networks

A vector of model parameters

A vector of sufficient statistics

$$\Pr(\mathbf{Y} = \mathbf{y} \mid \theta, \mathbf{X}) = \frac{\exp\{\theta^{\mathbf{t}} s(\mathbf{y}, \mathbf{X})\}}{\sum_{\mathbf{y}' \in \mathcal{Y}} \exp\{\theta^{\mathbf{t}} s(\mathbf{y}', \mathbf{X})\}}, \quad \forall \mathbf{y} \in \mathcal{Y}$$

Observed data

The normalizing constant

All possible networks

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- This allow us to directly compute **the normalizing constant**.
- Using the exact likelihood opens a huge window of methodological-possibilities.

ERGMs for Small Networks

A vector of model parameters

A vector of sufficient statistics

$$\Pr\left(\mathbf{Y} = \mathbf{y} \mid \theta, \mathbf{X}\right) = \frac{\exp\left\{\theta^{\mathbf{t}} \boldsymbol{s}\left(\mathbf{y}, \mathbf{X}\right)\right\}}{\sum_{\mathbf{y}' \in \mathcal{Y}} \exp\left\{\theta^{\mathbf{t}} \boldsymbol{s}\left(\mathbf{y}', \mathbf{X}\right)\right\}}, \quad \forall \mathbf{y} \in \mathcal{Y}$$

Observed data

The normalizing constant

All possible networks

- In the case of small-enough networks, computation of the likelihood becomes computationally feasible.
- This allow us to directly compute **the normalizing constant**.
- Using the exact likelihood opens a huge window of methodological-possibilities.
- We implemented this and more in the `ergmito` R package

Sidetrack...

ito, ita: From the latin -*ī̄ttus*. suffix in Spanish used to denote small or affection. e.g.:

*¡Qué lindo ese perr**ito**! / What a beautiful little dog!*

*¿Me darías una tac**ita** de azúcar? / Would you give me a small cup of sugar?*

Sidetrack...

ito, ita: From the latin -*ī̄ttus*. suffix in Spanish used to denote small or affection. e.g.:

*¡Qué lindo ese perr**ito**! / What a beautiful little dog!*

*¿Me darías una tac**ita** de azúcar? / Would you give me a small cup of sugar?*

Special thanks to George Barnett who proposed the name during the 2018 NASN!

The ergmito

In general

- Implements estimation of ERGMs using exact statistics for small networks.
- Meta-programming allows specifying likelihood (and gradient) functions for pooled models.
- Includes tools for simulating and post-estimation checks.
- Getting ready for CRAN!

The ergmito

In general

- Implements estimation of ERGMs using exact statistics for small networks.
- Meta-programming allows specifying likelihood (and gradient) functions for pooled models.
- Includes tools for simulating and post-estimation checks.
- Getting ready for CRAN!

Other features

- Vectorized calculation of sufficient statistics.
- Scales up nicely (hundreds of small networks) saving space and computation (when possible).
- Highly tested (90% coverage with more than one hundred tests).

ergmito example

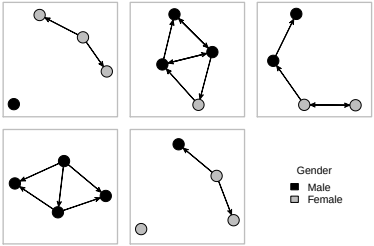


Figure 1 Random sample of 5 networks simulated using the ergmito package

ergmito example

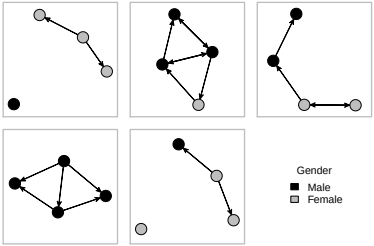


Figure 1 Random sample of 5 networks simulated using the ergmito package

	Bernoulli	Full model
Edge-count	-0.69^{*}	-1.70^{**}
	(0.27)	(0.54)
Homophily (on Gender)		1.59^{*}
		(0.64)
AIC	78.38	73.34
BIC	80.48	77.53
Log Likelihood	-38.19	-34.67
Num. networks	5	5

Standard errors in parenthesis. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1 Fitted ERGMitos using the fivenets dataset.

ergmito example

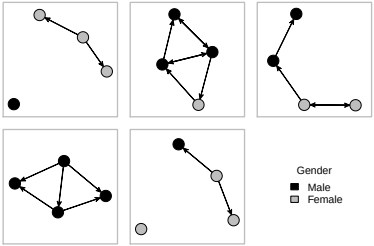


Figure 1 Random sample of 5 networks simulated using the ergmito package

	Bernoulli	Full model
Edge-count	-0.69^{*}	-1.70^{**}
	(0.27)	(0.54)
Homophily (on Gender)		1.59^{*}
		(0.64)
AIC	78.38	73.34
BIC	80.48	77.53
Log Likelihood	-38.19	-34.67
Num. networks	5	5

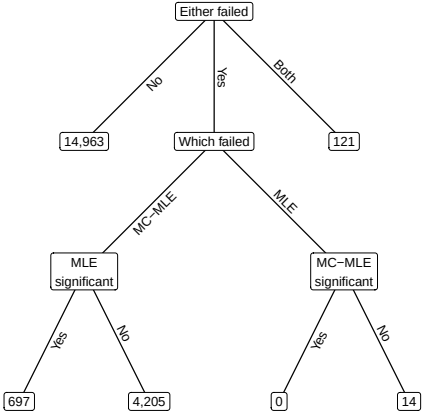
Standard errors in parenthesis. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1 Fitted ERGMitos using the fivenets dataset.

We performed a large simulation study more comparing MC-MLE (ergm) with MLE (ergmito).

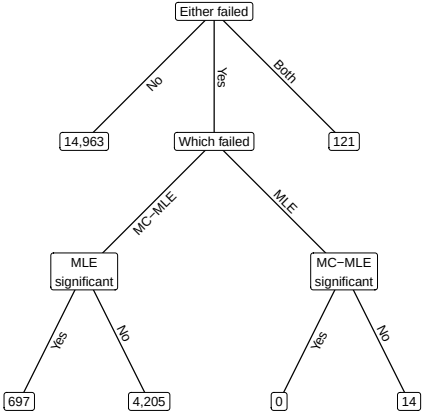
Paper 2 Simulation Studies: Error rate

- The MC-MLE implementation failed $\sim$ 5,000/20,000 times



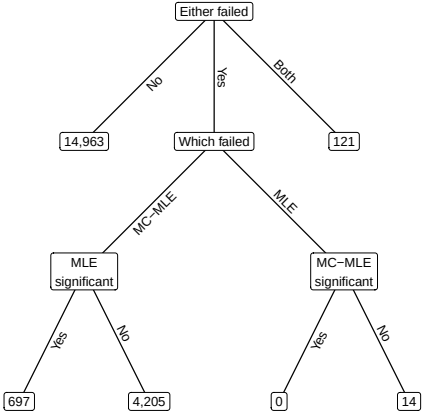
Paper 2 Simulation Studies: Error rate

- The MC-MLE implementation failed $\sim$ 5,000/20,000 times
- In $\sim$700 of those cases ergmito (MLE) reported a significant effect



Paper 2 Simulation Studies: Error rate

- The MC-MLE implementation failed $\sim$ 5,000/20,000 times
- In $\sim$700 of those cases ergmito (MLE) reported a significant effect
- I no case that MLE failed MC-MLE reported an effect.



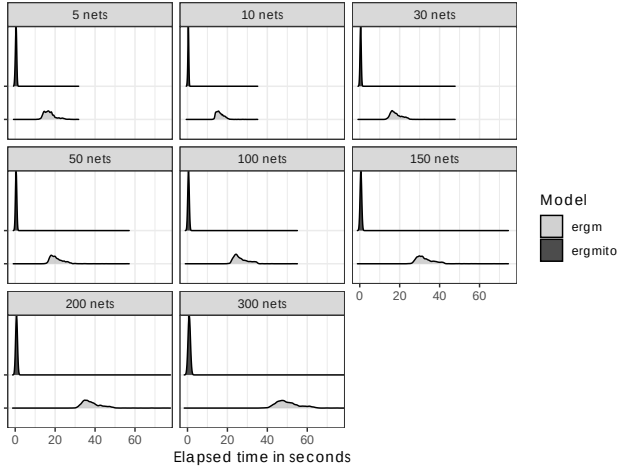
Paper 2 Simulation Studies: Empirical Type I error

Keck School of Medicine of USC

Sample size	P(Type I error)		χ^2
	MC-MLE (`ergm`)	MLE (`ergmito`)	
5	0.084	0.057	11.71 ***
10	0.070	0.045	12.46 ***
15	0.084	0.066	5.55 *
20	0.074	0.060	3.58
30	0.057	0.052	0.67
50	0.046	0.044	0.17
100	0.048	0.048	0.00

Table 2 Empirical Type I error rates. The χ^2 statistic is from a 2-sample test for equality of proportions, and the significance levels are given by *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$.

Paper 2 Simulation Studies: Elapsed time



more results

Keck School of Medicine of USC

Key takeaways

Paper 2: Exponential Random Graph Models for Small Networks

Keck School of Medicine of USC

Key takeaways

- New extension of ERGMs using exact statistics for small networks (families, teams, etc.)

Keck School of Medicine of USC

Key takeaways

- New extension of ERGMs using exact statistics for small networks (families, teams, etc.)
- Performance: Same (un)bias, Lower Type I error rates, (way) faster.

Paper 2: Exponential Random Graph Models for Small Networks

Keck School of Medicine of USC

Key takeaways

- New extension of ERGMs using exact statistics for small networks (families, teams, etc.)
- Performance: Same (un)bias, Lower Type I error rates, (way) faster.
- Opens the door the new methods, e.g. Mixed effects, LRT, etc.

Paper 2: Exponential Random Graph Models for Small Networks

Keck School of Medicine of USC

Key takeaways

- New extension of ERGMs using exact statistics for small networks (families, teams, etc.)
- Performance: Same (un)bias, Lower Type I error rates, (way) faster.
- Opens the door the new methods, e.g. Mixed effects, LRT, etc.

Challenges

- Computationally, we can do better in terms of speed/memory.

Paper 2: Exponential Random Graph Models for Small Networks

Keck School of Medicine of USC

Key takeaways

- New extension of ERGMs using exact statistics for small networks (families, teams, etc.)
- Performance: Same (un)bias, Lower Type I error rates, (way) faster.
- Opens the door the new methods, e.g. Mixed effects, LRT, etc.

Challenges

- Computationally, we can do better in terms of speed/memory.
- Have a good way of assessing goodness-of-fit.

Paper 2: Exponential Random Graph Models for Small Networks

Keck School of Medicine of USC

Key takeaways

- New extension of ERGMs using exact statistics for small networks (families, teams, etc.)
- Performance: Same (un)bias, Lower Type I error rates, (way) faster.
- Opens the door the new methods, e.g. Mixed effects, LRT, etc.

Challenges

- Computationally, we can do better in terms of speed/memory.
- Have a good way of assessing goodness-of-fit.
- Explore extending this method for (very) large networks.

Future Research

Keck School of Medicine of USC

Goodness-of-fit

Future Research: ERGMitos

Keck School of Medicine of USC

Goodness-of-fit

- Is something that will need to be addressed at some point.

Future Research: ERGMitos

Goodness-of-fit

- Is something that will need to be addressed at some point.
- The problem is not easy as we need to deal a discrete distribution.

Future Research: ERGMitos

Goodness-of-fit

- Is something that will need to be addressed at some point.
- The problem is not easy as we need to deal a discrete distribution.
- Two key questions: What sufficient statistic to look at? what test?

Future Research: ERGMitos

Keck School of Medicine of USC

Goodness-of-fit

- Is something that will need to be addressed at some point.
- The problem is not easy as we need to deal a discrete distribution.
- Two key questions: What sufficient statistic to look at? what test?

ERGMs for large networks

Future Research: ERGMitos

Keck School of Medicine of USC

Goodness-of-fit

- Is something that will need to be addressed at some point.
- The problem is not easy as we need to deal a discrete distribution.
- Two key questions: What sufficient statistic to look at? what test?

ERGMs for large networks

- There is still no standard way to estimate ERGMs for large networks.

Future Research: ERGMitos

Keck School of Medicine of USC

Goodness-of-fit

- Is something that will need to be addressed at some point.
- The problem is not easy as we need to deal a discrete distribution.
- Two key questions: What sufficient statistic to look at? what test?

ERGMs for large networks

- There is still no standard way to estimate ERGMs for large networks.
- Most attempts are still depending on simulation methods.

Future Research: ERGMitos

Keck School of Medicine of USC

Goodness-of-fit

- Is something that will need to be addressed at some point.
- The problem is not easy as we need to deal a discrete distribution.
- Two key questions: What sufficient statistic to look at? what test?

ERGMs for large networks

- There is still no standard way to estimate ERGMs for large networks.
- Most attempts are still depending on simulation methods.
- We could use the Snowball Sampling framework together with ERGMitos.

Future Research: ERGMitos

Keck School of Medicine of USC

Goodness-of-fit

- Is something that will need to be addressed at some point.
- The problem is not easy as we need to deal a discrete distribution.
- Two key questions: What sufficient statistic to look at? what test?

ERGMs for large networks

- There is still no standard way to estimate ERGMs for large networks.
- Most attempts are still depending on simulation methods.
- We could use the Snowball Sampling framework together with ERGMitos. (... I would call this ERGMote)

Concluding Remarks

- An extension to a well studied models for social networks.

Concluding Remarks

- An extension to a well studied models for social networks.
- Small size allows exact calculations.

Concluding Remarks

- An extension to a well studied models for social networks.
- Small size allows exact calculations.
- Opens the door to a large set of methodological innovations.

Concluding Remarks

- An extension to a well studied models for social networks.
- Small size allows exact calculations.
- Opens the door to a large set of methodological innovations.
- **Next steps:** GOF or extensions to large networks?

Concluding Remarks

Keck School of Medicine of USC

- An extension to a well studied models for social networks.
- Small size allows exact calculations.
- Opens the door to a large set of methodological innovations.
- **Next steps:** GOF or extensions to large networks?

Accomplishments during the development of this work

- 6 journal publications (Journal of Open Source Software, Stata Journal, Journal of health and social behavior, Translational behavioral medicine, Social Science & Medicine)
- 11 packages/libraries built (ergmito, similR, gnet, fmcmc, slurmR, aphylo, polygons, pruner, netplot, rphyloxml, jsPhyloSVG)

Essays on Bioinformatics and Social Network Analysis

Statistical and Computational Methods for Complex Systems

George G Vega Yon

University of Southern California, Department of Preventive Medicine

January 28, 2020

Keck School of Medicine of USC

Thanks!

References I

ERGMs: The MC-MLE approach

One of the most popular methods for estimating ERGMs is the MC-MLE approach (citations here)

This consists on the following steps

1. Start from a sensible guess on what should be the population parameters (usually done using pseudo-MLE estimation)
2. While the algorithm doesn't converge, do:
   - 2.1 Simulate a stream of networks with the current state of the parameter, θ_t
   - 2.2 Using the law of large numbers, approximate the ratio of likelihoods based on the parameter θ_t, this is the objective function
   - 2.3 Update the parameter by a Newton-Raphson step
   - 2.4 Next iteration

go back

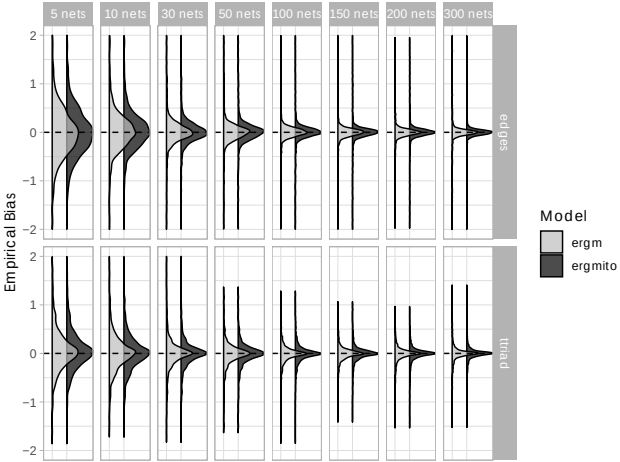
Paper 2 Simulation Studies

We performed a simulation study with the following features:

- Draw 20,000 samples of groups of small networks
- Each group had prescribed: (model parameters, number of networks, sizes of the networks)
- Each group could have from 5 to 300 small networks
- We estimated the models using MC-MLE and MLE.

go back

Paper 2 Simulation Studies: Empirical Bias



go back