Exact Statistics and Semi-Parametric Tests for Small Network Data

George G. Vega Yon, MS[⋆] Andrew Slaughter, PhD[†] Kayla de la Haye, PhD[⋆]

[⋆]University of Southern California
Department of Preventive Medicine

[†]U.S. Army Research Institute for the
Behavioral and Social Sciences

IC^2S^2 2019, Amsterdam
July 19, 2019

Acknowledgements

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 (hpc.usc.edu).

The views expressed in this presentation are those of the authors, and do not represent the official policy or positions of the Department of the Army, the DoD, or the U.S. Government.

Network Science of Teams

a Multidisciplinary University Research Initiative

UC SANTA BARBARA · USC University of Southern California · MIT Massachusetts Institute of Technology · NORTHWESTERN UNIVERSITY

Context: Social abilities and team performance

Two research questions

Context: Social abilities and team performance

Two research questions

How do **social abilities** impact **network structure**?

Context: Social abilities and team performance

Two research questions

How do **social abilities** impact **network structure**?

How does **collective intelligence** affect team (network) **performance**?

Context: Social abilities and team performance

Two research questions

How do **social abilities** impact **network structure**?

How does collective intelligence affect team (network) performance?

To answer this question, we have the following experimental data:

To answer this question, we have the following experimental data:

- 42 mixed-gender teams,

To answer this question, we have the following experimental data:

- 42 mixed-gender teams,
- Which completed 1 hour of group tasks (MIT's IQ test for teams)

To answer this question, we have the following experimental data:

- 42 mixed-gender teams,
- Which completed 1 hour of group tasks (MIT's IQ test for teams)
- Individual survey capturing information regarding socio-demographics **and**:

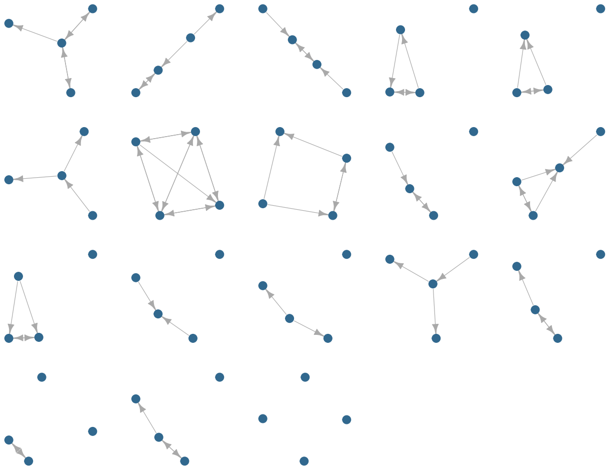
To answer this question, we have the following experimental data:

- 42 mixed-gender teams,
- Which completed 1 hour of group tasks (MIT's IQ test for teams)
- Individual survey capturing information regarding socio-demographics **and**:
  - **Social Intelligence**: Social Perception (measured by RME), Social Accomodation, Social Gregariousness, and Social Awareness

To answer this question, we have the following experimental data:

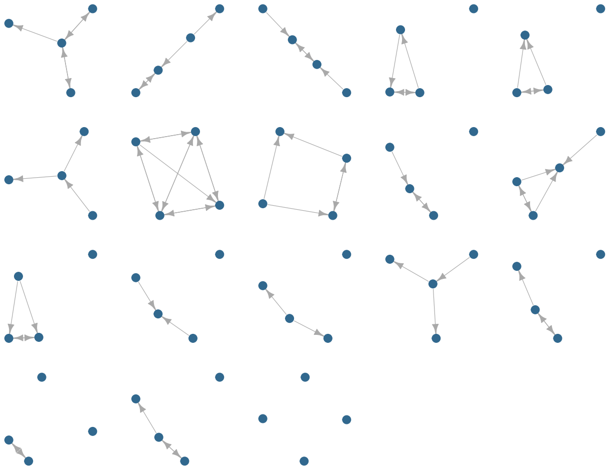
- 42 mixed-gender teams,
- Which completed 1 hour of group tasks (MIT's IQ test for teams)
- Individual survey capturing information regarding socio-demographics **and**:
  - **Social Intelligence**: Social Perception (measured by RME), Social Accomodation, Social Gregariousness, and Social Awareness
  - **Social Networks**: Advice Seeking, Leadership, Influence (among others).

Context (cont'd)



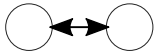
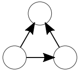
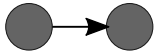
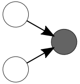
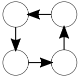
We can do a lot of simple statistics: density, % of *[blank]*, etc. but...

Context (cont'd)



We can do a lot of simple statistics: density, % of *[blank]*, etc. but... **how can we go beyond that?**

Exponential random graph models

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

ERGMs can do the job.

Exponential random graph models (a 1 slide crah course)

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

There is one problem with this model ...

A vector of model parameters

A vector of sufficient statistics

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

All possible networks

The normalizing constant

There is one problem with this model ...

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

because of $\mathcal{Y}$, the **normalizing constant** is

a summation of $2^{n(n-1)}$ terms !

There is one problem with this model ...

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

because of $\mathcal{Y}$, the **normalizing constant** is a summation of $2^{n(n-1)}$ terms 😱!

ERGMs for small networks

- Calculating the likelihood function for a directed graph means (at some point) enumerating $2^{n(n-1)}$ **terms**.

$$\Pr\left(\mathbf{G} = \mathbf{g} \mid \theta, \mathbf{X}\right) = \frac{\text{exp}\left\{\theta^{\mathbf{t}} s\left(\mathbf{g}, \mathbf{X}\right)\right\}}{\sum_{\mathbf{g}' \in \mathcal{G}} \text{exp}\left\{\theta^{\mathbf{t}} s\left(\mathbf{g}', \mathbf{X}\right)\right\}}$$

ERGMs for small networks

- Calculating the likelihood function for a directed graph means (at some point) enumerating $2^{n(n-1)}$ **terms**.

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

- So, if $n = 6$, then we have approx 1,000,000,000 terms.

ERGMs for small networks

- Calculating the likelihood function for a directed graph means (at some point) enumerating $2^{n(n-1)}$ **terms**.

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

- So, if $n=6$, then we have approx 1,000,000,000 terms.
- This has lead the field to aim for (very neat) simulation based methods

ERGMs for small networks

- Calculating the likelihood function for a directed graph means (at some point) enumerating $2^{n(n-1)}$ **terms**.

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

- So, if $n = 6$, then we have approx 1,000,000,000 terms.
- This has lead the field to aim for (very neat) simulation based methods
- But, if our small networks have (at most) 6 nodes...

ERGMs for small networks

- Calculating the likelihood function for a directed graph means (at some point) enumerating $2^{n(n-1)}$ **terms**.

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

- So, if $n = 6$, then we have approx 1,000,000,000 terms.
- This has lead the field to aim for (very neat) simulation based methods
- But, if our small networks have (at most) 6 nodes...

We can go back to the good-old-fashion MLE

Keeping $n \leq 6$ we can

Keeping $n \leq 6$ we can

- Compute the likelihood function exactly, and hence use ``simple'' optimization to get MLEs.

Keeping $n \leq 6$ we can

- Compute the likelihood function exactly, and hence use ``simple'' optimization to get MLEs.
- Obtain more **accurate** estimates **faster**

Keeping $n \leq 6$ we can

- Compute the likelihood function exactly, and hence use ``simple'' optimization to get MLEs.
- Obtain more **accurate** estimates **faster** (in most cases).

Keeping $n \leq 6$ we can

- Compute the likelihood function exactly, and hence use ``simple'' optimization to get MLEs.
- Obtain more **accurate** estimates **faster** (in most cases).
- Since (usually) small networks come in many...

Keeping $n \leq 6$ we can

- Compute the likelihood function exactly, and hence use ``simple'' optimization to get MLEs.
- Obtain more **accurate** estimates **faster** (in most cases).
- Since (usually) small networks come in many…obtain pooled estimates. Which helps with **power** *and* **degeneracy**)

Keeping $n \leq 6$ we can

- Compute the likelihood function exactly, and hence use ``simple'' optimization to get MLEs.
- Obtain more **accurate** estimates **faster** (in most cases).
- Since (usually) small networks come in many...obtain pooled estimates. Which helps with **power** *and* **degeneracy**)
- And more:
  - All MLE goodies, e.g., LRT
  - Enhanced simulation methods: resampling, cross-validation
  - Trivially extend ERGM: mixed-effects models, dependency structures across net

Keeping $n \leq 6$ we can

- Compute the likelihood function exactly, and hence use ``simple'' optimization to get MLEs.
- Obtain more **accurate** estimates **faster** (in most cases).
- Since (usually) small networks come in many...obtain pooled estimates. Which helps with **power** *and* **degeneracy**)
- And more:
  - All MLE goodies, e.g., LRT
  - Enhanced simulation methods: resampling, cross-validation
  - Trivially extend ERGM: mixed-effects models, dependency structures across net
  - etc.

Keeping $n \leq 6$ we can

- Compute the likelihood function exactly, and hence use ``simple'' optimization to get MLEs.
- Obtain more **accurate** estimates **faster** (in most cases).
- Since (usually) small networks come in many...obtain pooled estimates. Which helps with **power** *and* **degeneracy**)
- And more:
  - All MLE goodies, e.g., LRT
  - Enhanced simulation methods: resampling, cross-validation
  - Trivially extend ERGM: mixed-effects models, dependency structures across net
  - etc.

This and more has been implemented in the `ergmito` (lifecycle experimental) R package (available at https://github.com/muriteams/ergmito)

Sidetrack...

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

e.g.:

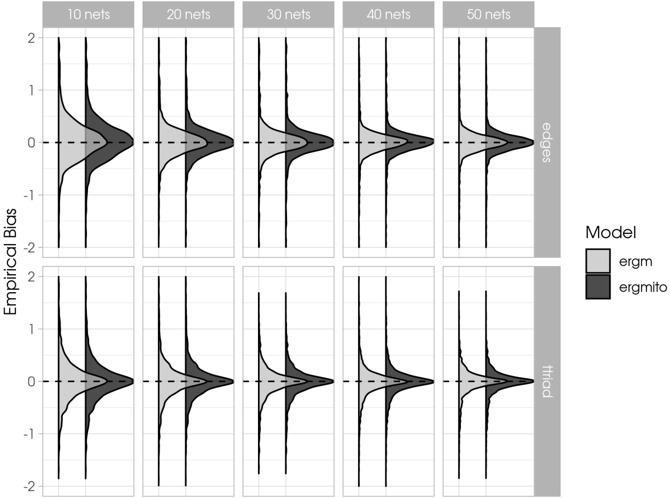
> *¡Qué lindo ese perr**ito**! / What a beautiful little dog!*
> *¿Me darías una tac**ita** de azúcar? / Would you give me a small cup of sugar?*

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

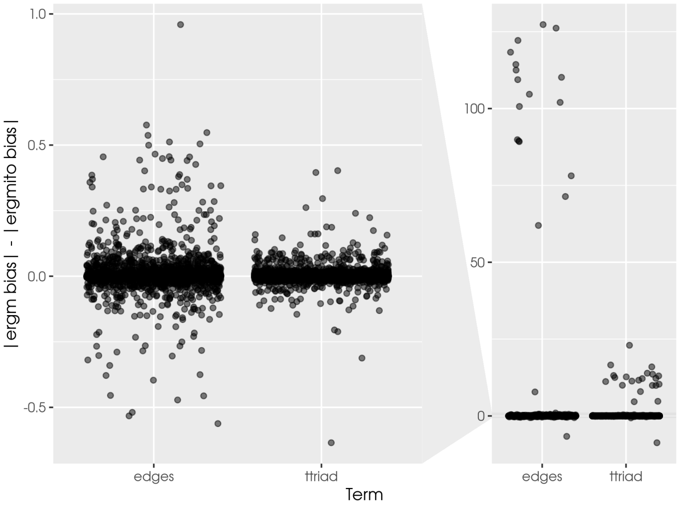
How many networks?

- Thinking about power and unbiasedness, we did a simulation study
- Simulated 20,000 samples of networks using the following steps:
  1. Draw parameters for the model based on the terms edges and ttriad (transitive triples) from a uniform(-2, 2).
  2. Draw group sizes by randomly selecting the number of networks of size 4 and size 5. Each sample has any of $\{10, 20, \dots, 50\}$ networks.
  3. Using 1. and 2., simulate networks using an ERGM model
- We fitted the models using both MC-MLE (block-diagonal ergm) and MLE (ergmito).
- We looked (are looking) at empirical bias (sanity check), power and elapsed time.

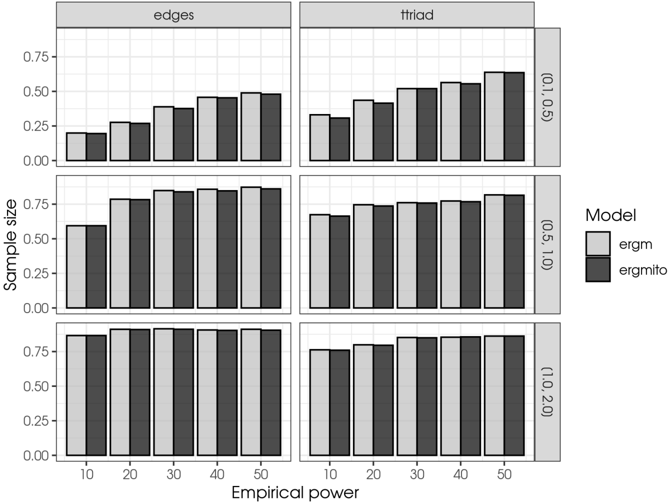
Simulation study: Bias



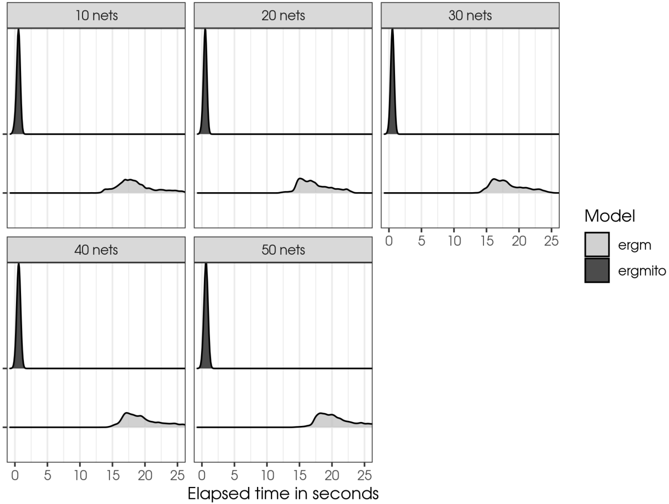
Simulation study: Bias (contd')



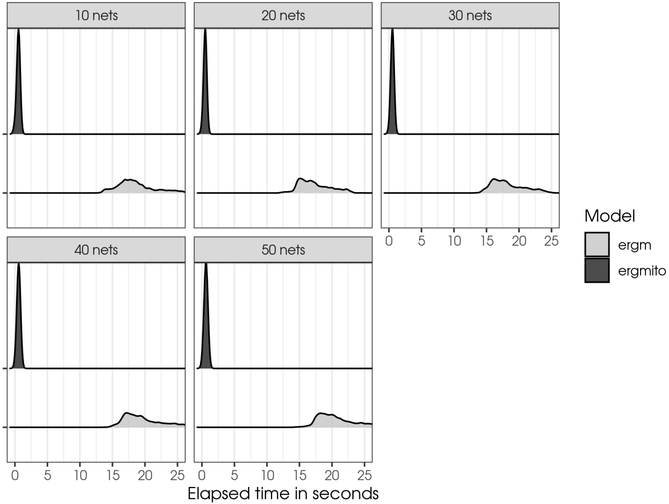
Simulation study: Power



Simulation study: Elapsed time (contd')



Simulation study: Elapsed time (contd')



What about a real data set?

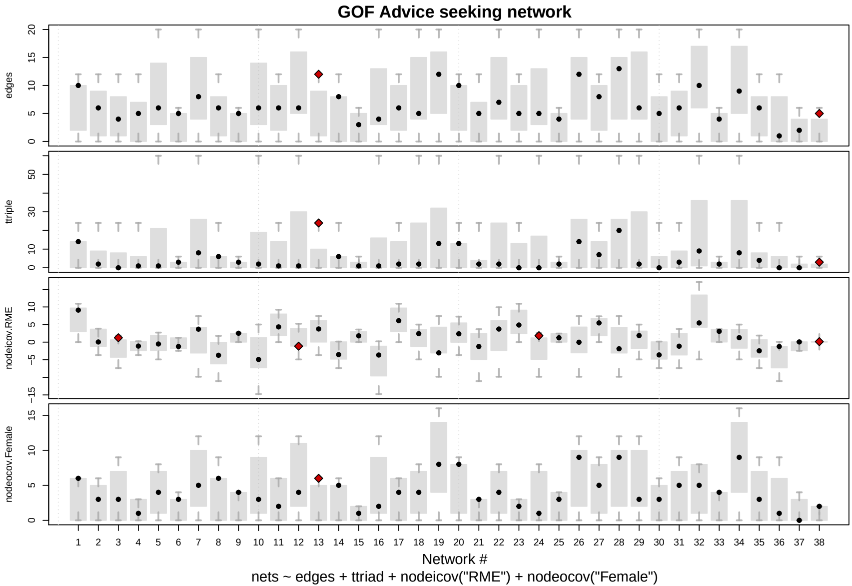
Preliminary results

From our sample of 42 small networks:

	Advice	Dislike	Influence	Leader	Trust
edges	-0.85^{***}	-2.30^{***}	-0.77^{***}	-0.53^{***}	-0.47^{***}
	(0.17)	(0.20)	(0.13)	(0.14)	(0.14)
ttriple	0.24^{***}		0.21^{**}		0.20^{***}
	(0.06)		(0.08)		(0.06)
nodeicov.RME	0.40^{***}		0.21^{*}	0.42^{***}	0.25^{**}
	(0.09)		(0.09)	(0.11)	(0.09)
nodeocov.Female	0.53^{**}				
	(0.18)				
nodematch.Female		0.56^{*}			
		(0.27)			
nodeicov.SI3Fac1		-0.35^{*}			
		(0.15)			
nodeicov.Female				-0.52^{**}	
				(0.20)	
nodeocov.RME				-0.32^{**}	
				(0.11)	
nodeocov.SI3Fac1					0.31^{***}
					(0.09)
AIC	695.07	381.72	756.84	637.01	776.82
BIC	712.13	394.52	769.92	654.07	794.25
Log Likelihood	-343.54	-187.86	-375.42	-314.50	-384.41
Num. networks	38	38	41	38	41
Convergence	0	0	0	0	0

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Selected models for each one of the studied networks. Results presented here correspond to a forward selection process.


GOF Advice seeking network
edges
ttriple
nodeicov.RME
nodeocov.Female
Network #
nets ~ edges + ttriad + nodeicov("RME") + nodeocov("Female")

Context: Social abilities and team performance

Two research questions

Context: Social abilities and team performance

Two research questions

~~How do **social abilities** impact **network structure**?~~

How does **collective intelligence** affect team (network) **performance**?

Networks and team performance

Suppose we have the following:

- Data on structure, nodes, and an outcome: $(\mathbf{g}, \mathbf{x}, y)$

Networks and team performance

Suppose we have the following:

- Data on structure, nodes, and an outcome: $(\mathbf{g}, \mathbf{x}, y)$
- In general, we are interested on assessing the following: $\mathbf{g} \perp y$

Networks and team performance

Suppose we have the following:

- Data on structure, nodes, and an outcome: $(\mathbf{g}, \mathbf{x}, y)$
- In general, we are interested on assessing the following: $\mathbf{g} \perp y$
- Three scenarios:

Networks and team performance

Suppose we have the following:

- Data on structure, nodes, and an outcome: $(\mathbf{g}, \mathbf{x}, y)$
- In general, we are interested on assessing the following: $\mathbf{g} \perp y$
- Three scenarios:
  a. Direct association

  b. Known (hypothesized) mediated association (a knwon confounder)

  c. Uknown (hypothesized) mediated association (an unknown confounder)

Networks and team performance

Suppose we have the following:

- Data on structure, nodes, and an outcome: $(\mathbf{g}, \mathbf{x}, y)$
- In general, we are interested on assessing the following: $\mathbf{g} \perp y$
- Three scenarios:

  a. Direct association

  b. Known (hypothesized) mediated association (a knwon confounder)

  c. Uknown (hypothesized) mediated association (an unknown confounder)

Non-parametrically, a. is ``trivial''

Networks and team performance

Suppose we have the following:

- Data on structure, nodes, and an outcome: $(\mathbf{g}, \mathbf{x}, y)$
- In general, we are interested on assessing the following: $\mathbf{g} \perp y$
- Three scenarios:
  - **a.** Direct association
  - **b.** Known (hypothesized) mediated association (a knwon confounder)
  - **c.** Uknown (hypothesized) mediated association (an unknown confounder)

Non-parametrically, a. is ``trivial'', c. is really hard,

Networks and team performance

Suppose we have the following:

- Data on structure, nodes, and an outcome: $(\mathbf{g}, \mathbf{x}, y)$
- In general, we are interested on assessing the following: $\mathbf{g} \perp y$
- Three scenarios:

  a. Direct association

  b. Known (hypothesized) mediated association (a knwon confounder)

  c. Uknown (hypothesized) mediated association (an unknown confounder)

Non-parametrically, a. is ``trivial'', c. is really hard, b. can be approached by:

a. Conditional permutation tests,

Networks and team performance

Suppose we have the following:

- Data on structure, nodes, and an outcome: $(\mathbf{g}, \mathbf{x}, y)$
- In general, we are interested on assessing the following: $\mathbf{g} \perp y$
- Three scenarios:

  a. Direct association

  b. Known (hypothesized) mediated association (a knwon confounder)

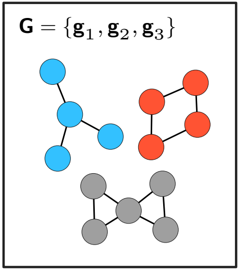
  c. Uknown (hypothesized) mediated association (an unknown confounder)

Non-parametrically, a. is ``trivial'', c. is really hard, b. can be approached by:

a. Conditional permutation tests,

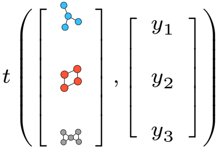
b. **Simulation based methods**

Step 1:
Fit the ERGMito



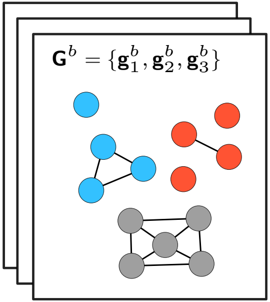
Fit the ERGMito,
This will give us $\mathcal{D}(\hat{\theta}, X_j)$

Step 2:
Calculate $t_0 =$



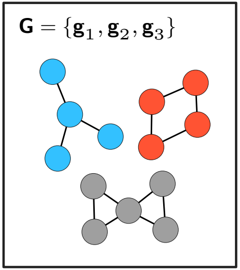
Throughout the simulations the only part that changes is the networks, not Y

Step 3:
For $b \in 1, \dots, B$ do



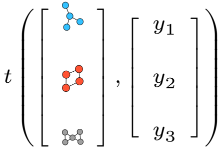
3.1) For $j \in \{1, 2, 3\}$ draw a new network from $\mathcal{D}$
3.2) Use the new sample to calculate $t_b = t(\mathbf{G}^b, Y)$

Step 1:
Fit the ERGMito



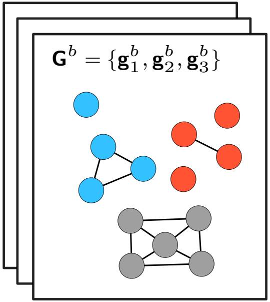
Fit the ERGMito,
This will give us $\mathcal{D}(\hat{\theta}, X_j)$

Step 2:
Calculate $t_0 =$



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

Step 3:
For $b \in 1, \dots, B$ do



3.1) For $j \in \{1, 2, 3\}$ draw a new network from $\mathcal{D}$
3.2) Use the new sample to calculate $t_b = t(\mathbf{G}^b, Y)$

We are still working (thinking) about this...

Discussion

- ERGMItos... This is not new.

Discussion

- ERGMItos... This is not new. What's new is the set of tools to apply it

Discussion

- ERGMItos... This is not new. What's new is the set of tools to apply it
- Taking this approach we can improve our estimates (power) and help with degeneracy

Discussion

- ERGMItos... This is not new. What's new is the set of tools to apply it
- Taking this approach we can improve our estimates (power) and help with degeneracy
- The tool is working

Discussion

- ERGMItos... This is not new. What's new is the set of tools to apply it
- Taking this approach we can improve our estimates (power) and help with degeneracy
- The tool is working (according to the simulation study...)

Discussion

- ERGMItos... This is not new. What's new is the set of tools to apply it
- Taking this approach we can improve our estimates (power) and help with degeneracy
- The tool is working (according to the simulation study...)
- Need to conduct more simulations using <u>nodal</u> attributes and larger samples of networks.

Discussion

- ERGMItos... This is not new. What's new is the set of tools to apply it
- Taking this approach we can improve our estimates (power) and help with degeneracy
- The tool is working (according to the simulation study...)
- Need to conduct more simulations using <u>nodal</u> attributes and larger samples of networks.
- What about goodness-of-fit? Still need to better think about it

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:
  - Separable Temporal ERGMitos, a.k.a. TERGMitos

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:
  - Separable Temporal ERGMitos, a.k.a. TERGMitos
  - Mixture models and Bayesian inference (if you are into that kind of stuff)

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:
  - Separable Temporal ERGMitos, a.k.a. TERGMitos
  - Mixture models and Bayesian inference (if you are into that kind of stuff)
  - More flexible formulas (e.g. interactions between terms and graph-level attributes)

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:
  - Separable Temporal ERGMitos, a.k.a. TERGMitos
  - Mixture models and Bayesian inference (if you are into that kind of stuff)
  - More flexible formulas (e.g. interactions between terms and graph-level attributes)
  - Better odds ratios (not simply exponentiating the coefficients)

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:
    - Separable Temporal ERGMitos, a.k.a. TERGMitos
    - Mixture models and Bayesian inference (if you are into that kind of stuff)
    - More flexible formulas (e.g. interactions between terms and graph-level attributes)
    - Better odds ratios (not simply exponentiating the coefficients)
    - Simulation based methods (small size $\implies$ sampling from in-memory data, and exact tests)

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:
  - Separable Temporal ERGMitos, a.k.a. TERGMitos
  - Mixture models and Bayesian inference (if you are into that kind of stuff)
  - More flexible formulas (e.g. interactions between terms and graph-level attributes)
  - Better odds ratios (not simply exponentiating the coefficients)
  - Simulation based methods (small size $\implies$ sampling from in-memory data, and exact tests)
  - Cross-validation/model selection in ERGMs

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:
  - Separable Temporal ERGMitos, a.k.a. TERGMitos
  - Mixture models and Bayesian inference (if you are into that kind of stuff)
  - More flexible formulas (e.g. interactions between terms and graph-level attributes)
  - Better odds ratios (not simply exponentiating the coefficients)
  - Simulation based methods (small size $\implies$ sampling from in-memory data, and exact tests)
  - Cross-validation/model selection in ERGMs
  - Fit models to large samples of small networks (10s of 1,000 vertices)

Discussion (contd')

- The simplicity of the estimation procedure allows us to think of:
  - Separable Temporal ERGMitos, a.k.a. TERGMitos
  - Mixture models and Bayesian inference (if you are into that kind of stuff)
  - More flexible formulas (e.g. interactions between terms and graph-level attributes)
  - Better odds ratios (not simply exponentiating the coefficients)
  - Simulation based methods (small size $\implies$ sampling from in-memory data, and exact tests)
  - Cross-validation/model selection in ERGMs
  - Fit models to large samples of small networks (10s of 1,000 vertices)
- Still thinking about how to test for association between network structure and group outcome

Thanks!

We thank members of our MURI research team, USC's Center for Applied Network Analysis, Garry Robins, Carter Butts, Johan Koskinen, Noshir Contractor, and attendees of the NASN 2018 conference for their comments.

George G. Vega Yon, MS Andrew Slaughter, PhD Kayla de la Haye, PhD

vegayon@usc.edu

https://ggvy.cl

gvegayon gvegayon