Power and multicollinearity in small networks: A discussion of

“Tale of Two Datasets: Representativeness and Generalisability of Inference for Samples of Networks”

JSM 2023
Toronto, Canada

George G. Vega Yon, Ph.D.
The University of Utah

2023-08-09

Overview

Highlights Krivitsky, Coletti, and Hens (2022)

What I highlight in their paper:

- Start to finish framework for multi-ERG models.
- Dealing with heterogeneous samples.
- Model building process.
- Goodness-of-fit analyses.

Two important missing pieces (for the next paper): power analysis and how to deal with collinearity in small networks.

Power analysis in ERGMs

Sample size in ERGMs

Two different questions: *How many nodes?* and "*How many networks?*"

Number of nodes (the usual question)

- Is the network bounded?
- If it is bounded, can we collect all the nodes?
- If we cannot collect all the nodes, can we do inference (Schweinberger, Krivitsky, and Butts 2017; Schweinberger et al. 2020)?

Number of networks (not so usual)

- There is a growing number of studies featuring multiple networks (e.g., egocentric studies).
- There's no clear way to do power analysis in ERGMs.
- In funding justification, power analysis is fundamental, so we need that.

A possible approach

We can leverage conditional ERG models for power analysis.

- Conditioning on one sufficient statistic results in a distribution invariant to the associated parameter, formally:

$$
\begin{aligned}
\mathrm{Pr}_{\mathcal{Y},\boldsymbol{\theta}}\left(\boldsymbol{Y}=\boldsymbol{y} \mid \boldsymbol{g}(\boldsymbol{y})_l = s_l\right) &= \frac{\mathrm{Pr}_{\mathcal{Y},\boldsymbol{\theta}}\left(\boldsymbol{g}(\boldsymbol{Y})_{-l} = \boldsymbol{g}(\boldsymbol{y})_{-l}, \boldsymbol{g}(\boldsymbol{y})_l = s_l\right)}{\sum_{\boldsymbol{y}'\in\mathcal{Y}:\boldsymbol{g}(\boldsymbol{y}')_l = s_l} \mathrm{Pr}_{\mathcal{Y},\boldsymbol{\theta}}\left(\boldsymbol{g}\left(\boldsymbol{Y}\right) = \boldsymbol{y}'\right)} \\
&= \frac{\exp\left\{{\boldsymbol{\theta}_{-l}}^{\boldsymbol{t}}\boldsymbol{g}(\boldsymbol{y})_{-l}\right\}}{\kappa_{\mathcal{Y}}(\boldsymbol{\theta})_{-l}},
\end{aligned} \tag{1}
$$

  where $\boldsymbol{g}(\boldsymbol{y})_l$ and $\boldsymbol{\theta}_l$ are the l-th element of $\boldsymbol{g}\left(\boldsymbol{y}\right)$ and $\boldsymbol{\theta}$ respectively, $\boldsymbol{g}(\boldsymbol{y})_{-l}$ and $\boldsymbol{\theta}_{-l}$ are their complement, and $\kappa_{\mathcal{Y}}(\boldsymbol{\theta})_{-l} = \sum_{\boldsymbol{y}'\in\mathcal{Y}:\boldsymbol{g}\left(\boldsymbol{y}'\right)_l = s_l} \exp\left\{{\boldsymbol{\theta}_{-l}}^{\boldsymbol{t}}\boldsymbol{g}\left(\boldsymbol{y}'\right)_{-l}\right\}$ is the normalizing constant.
- We can use this to generate networks with a prescribed edgecount (based on previous studies) and compute power through simulation.

Example: Detecting gender homophily

Want to **detect an effect size of** $\theta_{\text{homophily}} = 2$, using conditional ERGMs (prev Eq.):

1. For each $n \in N \equiv \{10, 20, \dots\}$, do:
   a. **Simulate**: $1,000$ sets of n undirected networks of size 8 and 26 ties.
   b. **Fit ERGM** Estimate $\hat{\theta}_{\text{homophily}}$, and generate the indicator variable $p_{n,i}$ equal to one if the estimate is significant at the 95% level.
   c. **Compute empirical power** $p_n \equiv \frac{1}{1{,}000}\sum_i p_{n,i}$.
2. **Model n as a function of power** Using $\{p_{10}, p_{20}, \dots\}$, we can fit the model $n \sim f(p_n)$.

Using KCH as a reference for density, we can fix the edge count to $0.93 \times 8(8-1)/2 \approx 26$

Parameter	Value
Network size	8
Edge count	26
$\theta_{\text{homophily}}$	2
α	0.10
$1-\beta$	0.80

Finally, the required sample size can be computed with $f(1-\beta) = f(0.80)$.

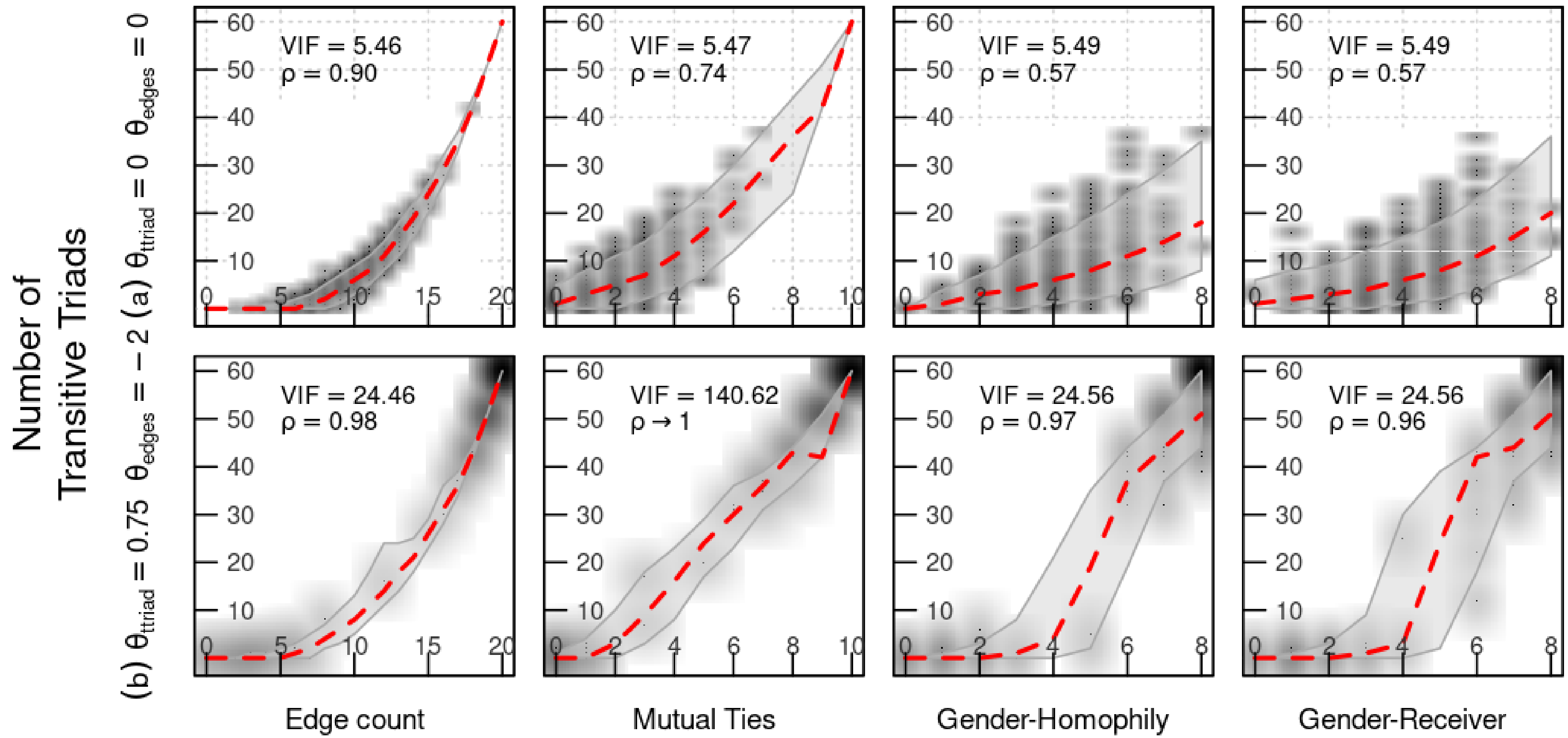
Collinearity in ERGMs

Not like in regular models

- Variance Inflation Factor [VIF] is a common measure of collinearity in regular models.
- Usually, VIF > 10 is considered problematic.
- VIFs are not straightforward in ERGMs:
  - Traditional models can feature completely exogenous variables.
  - ERGMs are by construction endogenous (**highly correlated**).
  - It is expected that VIFs will be higher in ERGMs.
- Duxbury (2021)'s large simulation study recommends using VIF between 20 and 150 as a threshold for multicollinearity.
- As small networks usually are denser, VIFs can be more severe.

Predicting statistics

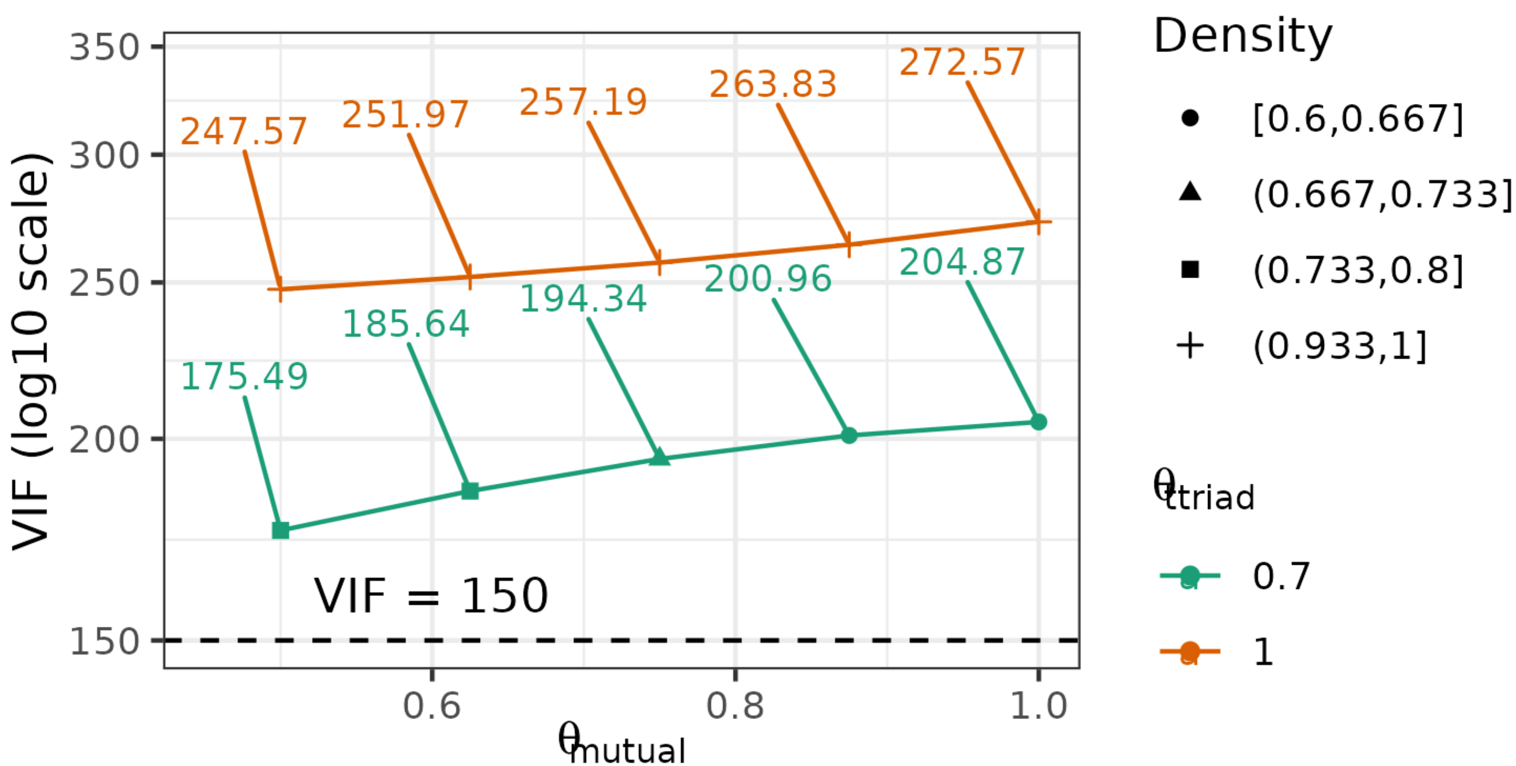
- A directed network with 5 nodes, two of them female and three male.
- Two models: **(a) Bernoulli** (0.50 density) and **(b) ERGM(edge count, transitivity)** (0.92 density).
- When $\theta_{\text{ttriad}} = 0.75$ and $\theta_{\text{edges}} = -2$ (second row), Cor(transitive triads, mutual ties) $\to 1$, and VIF reaches 140 (mutual ties).



Collinearity in small networks

- In the same network, many combinations of model parameters yield $\rho \to 1$ and high VIFs.
- KCH's networks were highly dense, (0.93 and 0.73 for the household and egocentric samples, respectively.) → collinearity should be severe.

$Y \sim \text{ERGM}(\text{edgecount, mutual ties, transitivity})$



Discussion

- Krivitsky, Coletti, and Hens' work make an important contribution to ERG models, most relevant: model building, selection, and GOF for multi-network models.
- Power (sample size requirements) and multicollinearity are two important issues that are yet to be addressed.
- I presented a possible approach to deal with power analysis in ERGMs using conditional distributions.
- Collinearity in small networks (like those in KCH) can be serious (more than in larger networks.) Yet we need to further explore this.

Thanks!

george.vegayon at utah.edu

https://ggv.cl

@gvegayon@qoto.org

References

Duxbury, Scott W. 2021. "Diagnosing Multicollinearity in Exponential Random Graph Models." *Sociological Methods & Research* 50 (2): 491–530. https://doi.org/10.1177/0049124118782543.

Krivitsky, Pavel N., Pietro Coletti, and Niel Hens. 2022. "A Tale of Two Datasets: Representativeness and Generalisability of Inference for Samples of Networks."

Schweinberger, Michael, Pavel N. Krivitsky, and Carter T. Butts. 2017. "A Note on the Role of Projectivity in Likelihood-Based Inference for Random Graph Models," July, 1–6.

Schweinberger, Michael, Pavel N. Krivitsky, Carter T. Butts, and Jonathan R. Stewart. 2020. "Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios." *Statistical Science* 35 (4): 627–62. https://doi.org/10.1214/19-sts743.