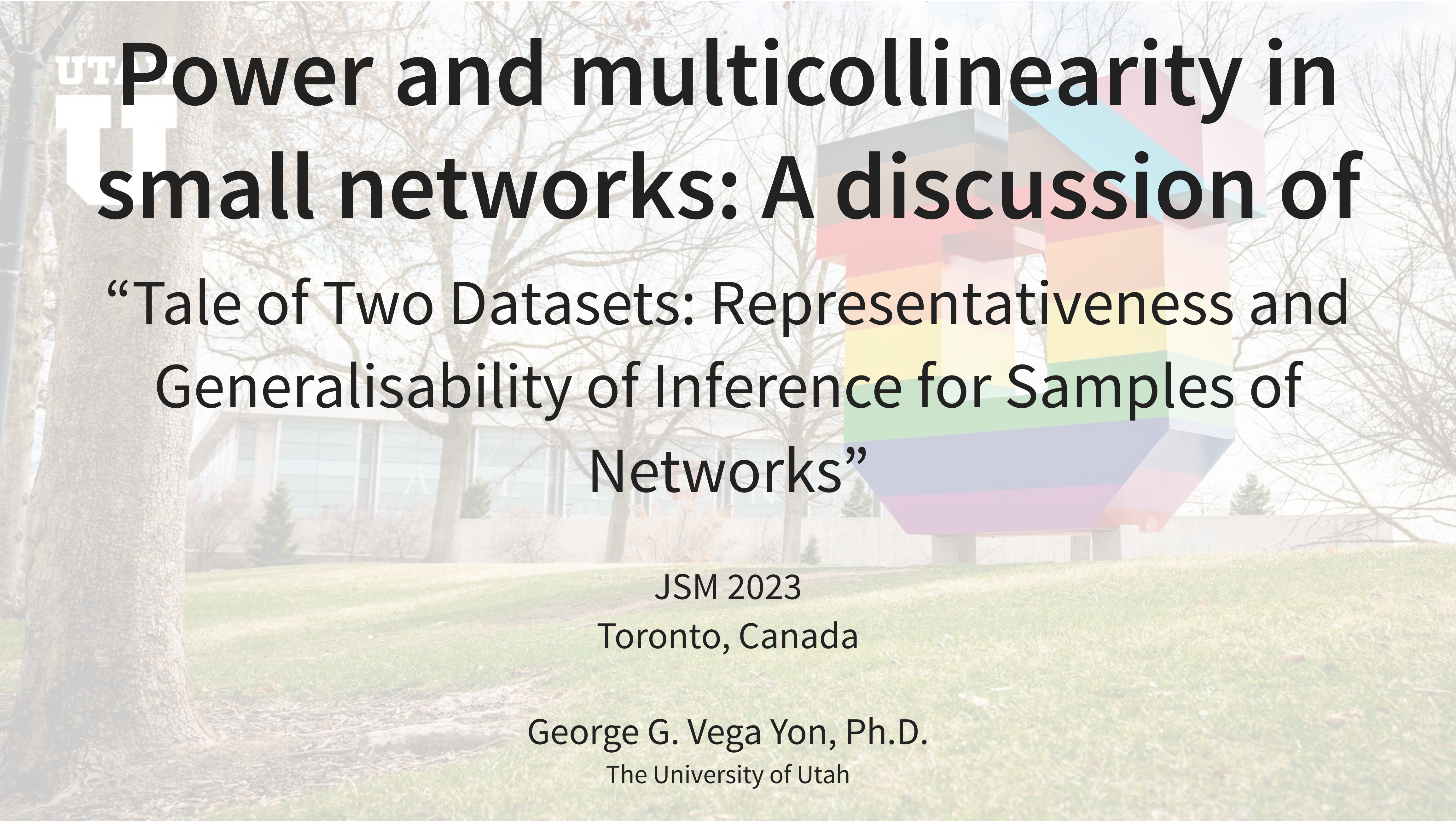


Power and multicollinearity in small networks: A discussion of

“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

2023-06-28

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

- 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

- 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:

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

  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 density (based on previous studies) and compute power through simulation.

Example: Detecting gender homophily

- Study gender homophily in networks of size 8.
- On average, the focal networks have 20 ties (, a density of $(2 \times 20)/(8 \times 7) \approx 0.71$).
- Want to detect an effect size of $\boldsymbol{\theta}_{\text{homophily}} = 2$, we could approximate the required sample size in the following fashion:

1. For each $n \in N \equiv \{10, 20, \dots\}$, do:

a. With Eq. (1), use MCMC to simulate $1,000$ sets of n networks of size 8 and 20 ties.

b. For each set, fit a conditional ERGM to estimate $\widehat{\boldsymbol{\theta}}_{\text{homophily}}$, and generate the indicator variable $p_{n,i}$ equal to one if the estimate is significant at the 95% level.

c. The empirical power for n is equal to $p_n \equiv \frac{1}{1{,}000} \sum_i p_{n,i}$.

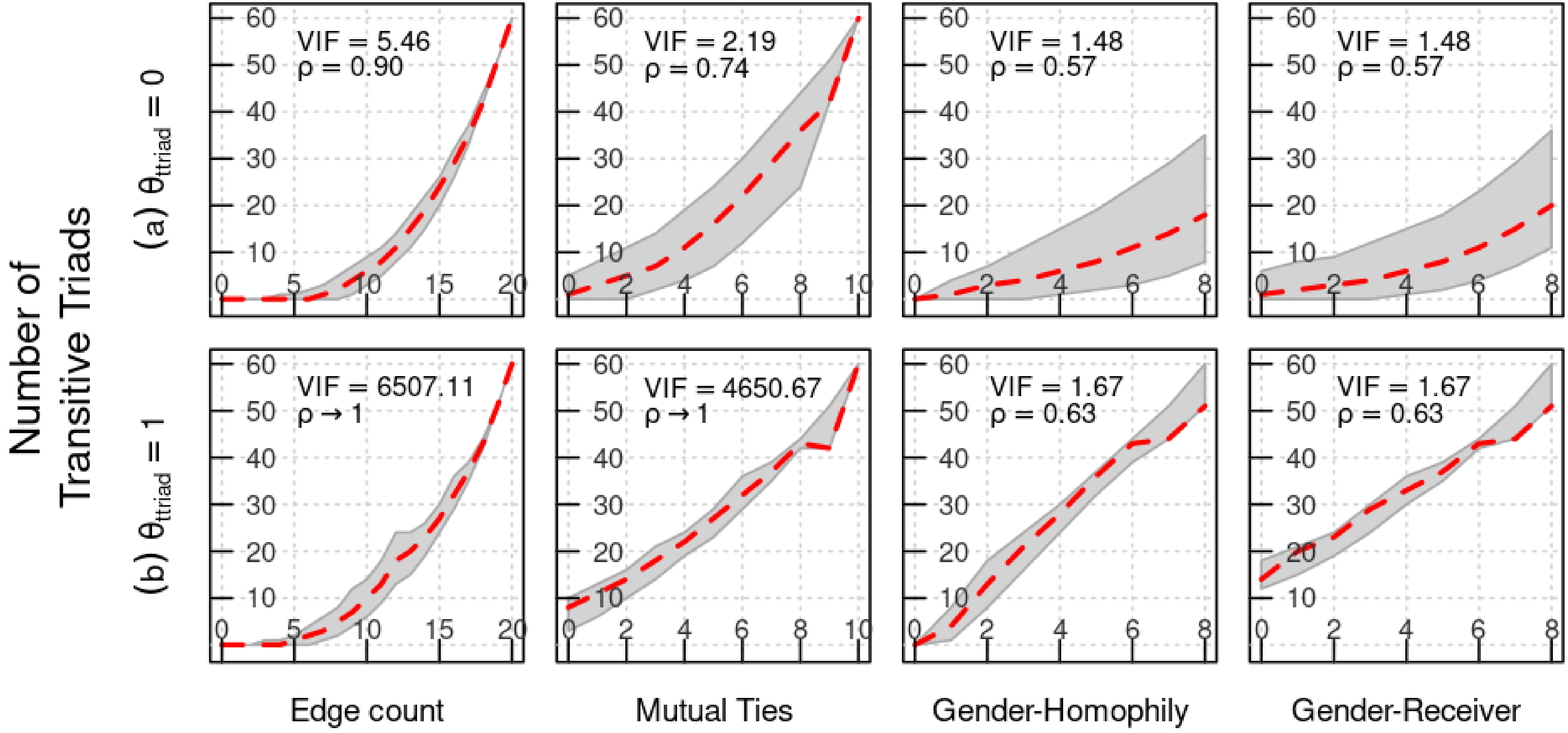
2. Once we have computed the sequence $\{p_{10}, p_{20}, \dots\}$, we can fit a linear model to estimate the sample size as a function of the power, , $n = \beta_0 + \beta_1 p_n + \beta_2 p_n^2 + \varepsilon$.
3. With the previous model in hand, we can estimate the sample size required to detect a given effect size with a given power.

Collinearity in ERGMs

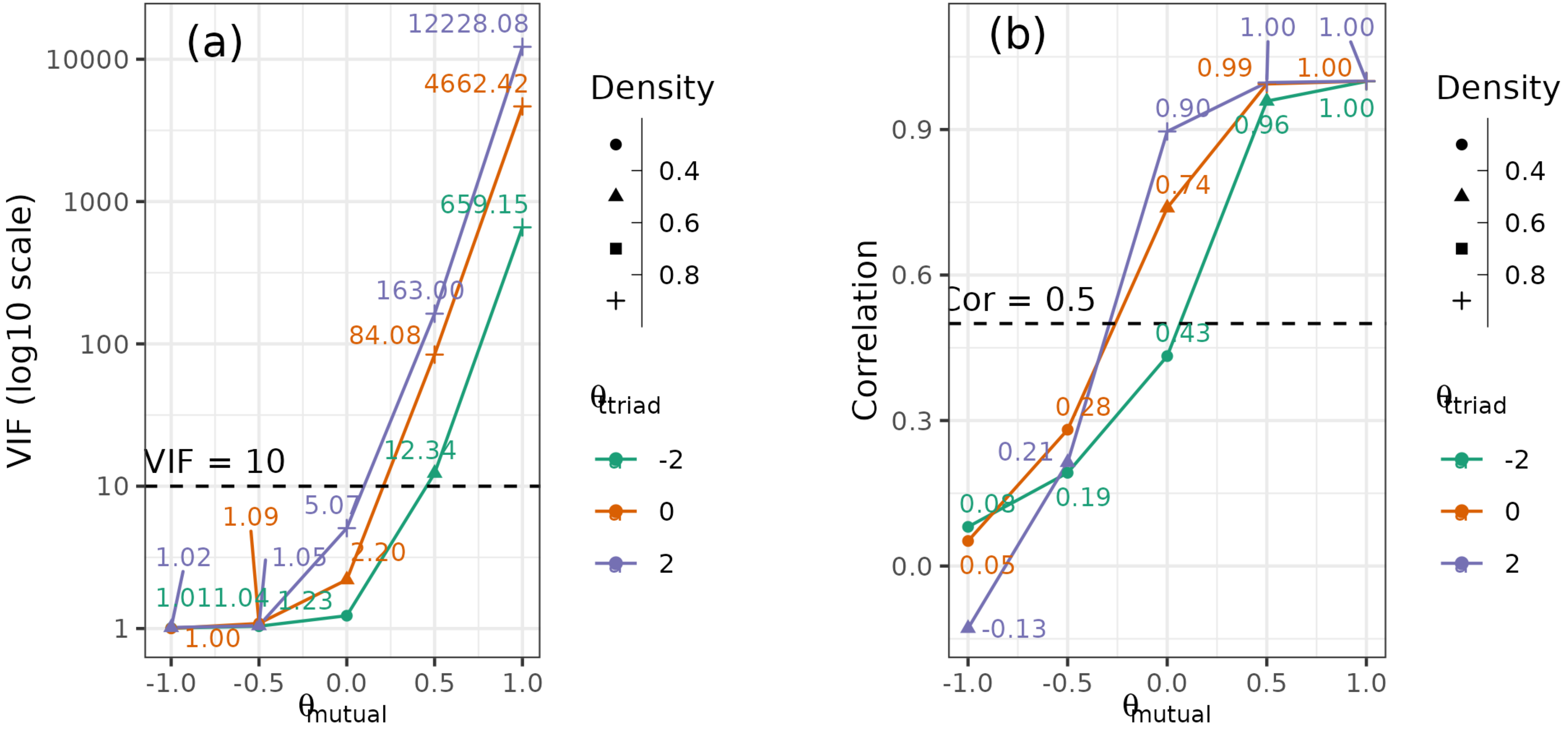
Not like in regular models

- Variance Inflation Factor [VIF] is a common measure of collinearity in regular models.
- Usually, VIF > 10 is considered problematic.
- Duxbury (2021)'s large simulation study recommends using VIF between 150 and 200 as a threshold for multicollinearity.
- In small networks, this could be more severe.

Predicting statistics



Collinearity in small networks



Discussion

A few questions:

- How would you address power analysis in ERGMs?
- VIFs and correlations across statistics are significantly high in dense networks. How much do you think it matters? If it matters, how would you address it?
- Relating both, is there any way in which a large sample size can help with collinearity?
- In KCH, effect sizes are significantly large.
- How much heterogeneity? Networks in KCH range from two to eight, but how about bigger samples? In Schweinberger, Krivitsky, and Butts (2017) it is mentioned the term “comparative”, but there’s no clear definition of what that means.

Thanks!

george.vegayon at utah.edu

https://ggv.cl

@gvegayon

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.