# Power and multicollinearity in small networks: A discussion of " Tale of Two Datasets: Representativeness and Generalisability of Inference for Samples of Networks"

George G. Vega Yon
Division of Epidemiology, University of Utah

July 20, 2023

## Abstract

The text of your abstract. 200 or fewer words.

*Keywords:* 3 to 6 keywords, that do not appear in the title

# 1 Introduction

The recent work by Krivitsky, Coletti & Hens provides an important new contribution to ~~the~~ Exponential-Family Random Graph [ERG] Models, a start-to-finish approach to dealing with multi-network ERGMs. Although multi-network ERGMs have been around for a while (mostly in the form of block-diagonal models, see Duxbury & Wertsching (2023)), not much care has been given to the estimation and post-estimation steps. In their paper, Krivitsky, Coletti & Hens give a detailed layout of how to build, estimate, and analyze multi-ERG models with heterogeneous data sources. Although their paper greatly contributes to inference and goodness of fit assessment, many issues still need to be resolved. In this comment, I will focus on two issues the authors did not discuss: power and multicollinearity, the latter regarding small networks.

# 2 Power in ERGMs: How many nodes/networks do we need?

Samples of multiple networks are becoming increasingly common (Vega Yon et al. 2021, Krivitsky et al. 2022, Duxbury & Wertsching 2023). Because of the complexity of ERGMs, power analysis (as in the required sample size for doing inference) has not been discussed in much of the literature other than in brief mentions (as in Vega Yon et al. (2021)). In this scenario, it is natural to start thinking about power analysis in ERGMs, especially in the case of small networks like those featured in egocentric studies, as, from the economic resources point of view, they are easier to collect. Power analysis ~~here is a bit more complex as we have two size parameters to pick~~: the network size and the number of networks.

**Network size**. Size for network studies, defined as the number of nodes to include in an ERG model, has two sides: network boundaries (as “finite” vs “continuous” in Butts (2009)) and statistical inference (as in projectivity). Many times the studied network has a well-defined boundary, for example, team networks and school friends. In such a scenario, surveying the entire network is completely feasible, so the question “How many nodes to survey?” is irrelevant. On the other hand, when the network boundaries are not well defined (like in a citation network), or the number of nodes in the network is large, “How many nodes are needed?” for statistical inference becomes relevant. In oversimplified terms, projectivity in our context refers to the validity of inference on a subset of the complete graph. Although not all ERGMs are projective, it has been shown that projectivity is not a necessary condition for consistency in likelihood-based inference (Schweinberger et al. 2017, 2020).

**Number of networks**. Here, in the context of multi-ERG models, the number of networks is akin to sample size in traditional statistical analyses. From the practical point of view, researchers’ capacity to generate data as that featured in Krivitsky, Coletti & Hens’s is tightly related to funding opportunities. Having a way of justifying the number of networks required in research projects is highly relevant.

As high computational power is increasingly accessible, one possible approach is to design a framework for power analysis in pooled ERGMs via simulation studies. For example, scientists exploring the role of gender in small teams could use observed data on density to estimate the number of networks needed to detect a given effect size using conditional ERGMs. Using the same notation as in Krivitsky et al. (2022), the conditional ERG probability mass function [pmf] given the $l$-th statistic, $\boldsymbol{g}\left(\boldsymbol{y}\right)_l = s_l$, can be calculated as follows:

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

where $\boldsymbol{g}\left(\boldsymbol{y}\right)_{l}$ and $\boldsymbol{\theta}_{l}$ are the $l$-th element of $\boldsymbol{g}\left(\boldsymbol{y}\right)$ and $\boldsymbol{\theta}$ respectively, $\boldsymbol{g}\left(\boldsymbol{y}\right)_{-l}$ and $\boldsymbol{\theta}_{-l}$ are their complement, and $\kappa_{\mathcal{Y}}\left(\boldsymbol{\theta}\right)_{-l}=\sum_{\boldsymbol{y}'\in\mathcal{Y}:\boldsymbol{g}\left(\boldsymbol{y}'\right)_{l}=s_{l}}\exp\left\{{\boldsymbol{\theta}_{-l}}^{\boldsymbol{t}}\boldsymbol{g}\left(\boldsymbol{y}'\right)_{-l}\right\}$ is the normalizing constant. In other words, once we condition on the $l$-th sufficient statistic, the *pmf* becomes invariant to the value of $\boldsymbol{\theta}_{l}$. Using the conditional distribution, we could simulate networks with a given density and then calculate the power of detecting gender homophily without needing to observe $\boldsymbol{\theta}_{\text{density}}$.

## 3 Collinearity in small networks

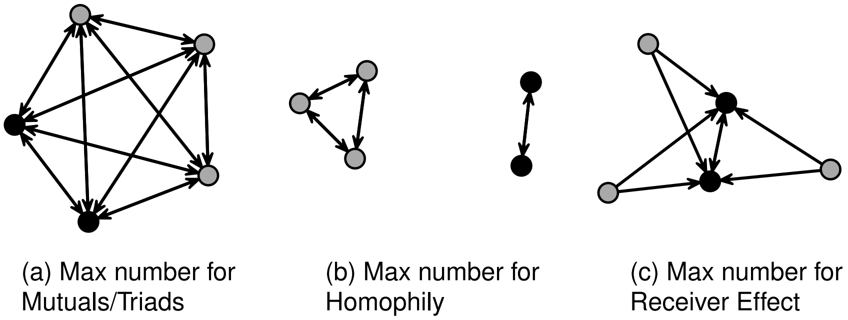
Since most of the statistics used in ERG models are by construction ~~function~~ of the number of ties in the network, the problem of collinearity is embedded in ERGMs. If we learned one thing from Duxbury (2021)'s large simulation study is that traditional measurements of collinearity are completely off in network models, with the Variance Inflation Factor [VIF] surpassing the rule of thumb of 10 by orders of magnitude ~~y~~ most cases.

As a proof of concept, I have selected four statistics that are commonly used in the ERGM literature, including the *Edge count* term, to analyze how these are interrelated using (1); these statistics are (a) *Number of Mutual Ties*, $\sum_{i\neq j}y_{ij}y_{ji}$; (b) *Number of Transitive Triads* (transitivity), $\sum_{i\neq j\neq k}y_{ij}y_{jk}y_{ik}$; (c) *Number of Homophilic Ties*, $\sum_{i\neq j}y_{ij}\mathbf{1}\left(x_i=x_j\right)$; (d) and *Attribute-Receiver effect*, $\sum_{i\neq j}y_{ij}x_j$, with the latter two using the binary attribute *Gender*, (which equals to one if the corresponding node is female and zero otherwise). Of these, only the *Number of Transitive Triads* and *Number of Mutual Ties* are Markovian in the sense of Frank & Strauss (1986). All calculations and figures were done using R version 4.3.1 (R Core Team 2023), the R packages **ergm** (Hunter et al. 2008, Handcock et al. 2023, Krivitsky et al. 2023), **ergmito** (Vega Yon 2023, Vega Yon et al. 2021), and **ggplot2** (Wickham 2016).

**Predictive power**. Figure 2 depicts the 95% Confidence Interval [CI] (shaded area), and fiftieth-percentile (red dashed-lines), of the number of transitive triads (y-axis) when conditioning on the others (x-axis), while at the same time, either assuming that the number of transitive triads has no say in the data-generating-process [DGP], (so fixing $\boldsymbol{\theta}_{\text{ttriad}} = 0$), four left figures, or on the contrary, assuming that the term is part of the DGP, (fixing $\boldsymbol{\theta}_{\text{ttriad}} = 1$), four right figures. The baseline graph used to calculate these statistics was a directed network of size five with the gender attribute equal to $(0, 0, 0, 1, 1)$, *i.e.*, three males and two females. Figure 1 illustrates three possible configurations, particularly the required arrangement (number and assignment of ties) needed for the analyzed sufficient statistics to reach their maximum values.

As seen in Figure 2, the CI for the expected number of transitive triads is generally very narrow. Since this statistic reflects relatively high-order structures (because more ties are involved), its range of possible values given the number of edge counts or mutual ties is highly constrained. Neither *Gender-Homophily* nor *Gender-Receiver Effect* seem to be very related to transitivity; yet, as $\boldsymbol{\theta}_{\text{ttriad}} : 0 \rightarrow 1$, all four conditioning statistics become very good at predicting the number of transitive triads in the graph. Furthermore, the VIF of transitivity as a function of mutual ties is almost 5,000 and the correlation between those two is almost 1.0. This wouldn't pass the recommendations made by Duxbury, who suggests that VIFs above 150 should be concerning.

Concerning the relatively poor predictive power of the *Gender-Homophily* and *Gender-*



(a) Max number for Mutuals/Triads

(b) Max number for Homophily

(c) Max number for Receiver Effect

Figure 1: **Networks with maximum observed counts.** All three networks have five vertices composed of two females (black vertices) and three males (gray vertices). Network (a) shows the required set of ties to reach the maximum value of *Mutuals* and *Transitive Triads*, which coincides with a fully connected graph, containing $(5 \times 4)/2 = 10$ mutual ties and $5 \times 4 \times 3 = 60$ transitive triads. Graphs (b) and (c) show the required configuration to maximize the number of *Gender-Homophily* and *Gender-Receiver Effect*, each equal to $3 \times 2 + 2 \times 1 = 8$ and $4 + 4 = 8$, respectively.

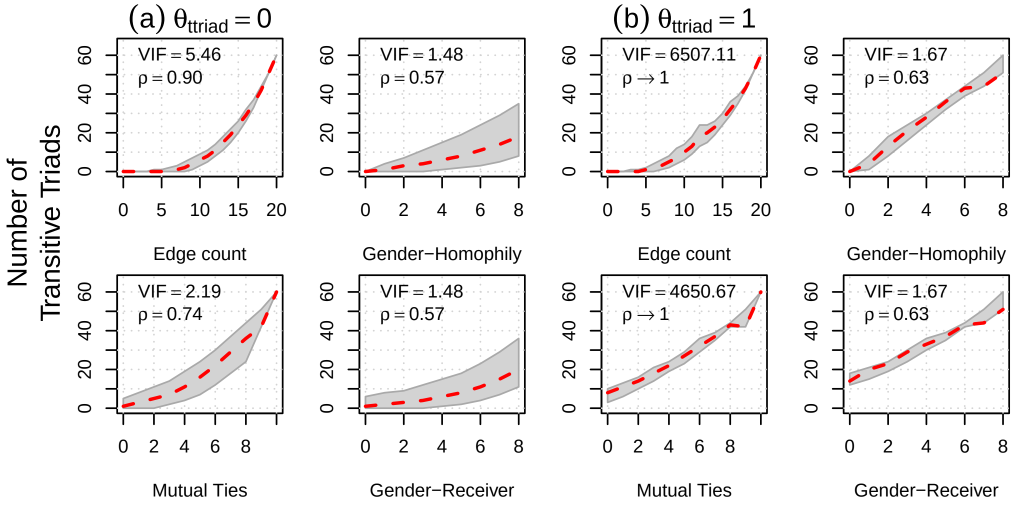


Figure 2: Conditional distribution of the number of transitive triads sufficient statistic. Column **(a)** shows the distribution under the Bernoulli model; this is, the parameter of transitivity is set to zero, whereas column **(b)** shows the distribution of the statistic assuming $\boldsymbol{\theta}_{ttriad} = 1$.

*Receiver Effect* sufficient statistics, besides the fact that these are non-Markovian, part of this could be because they reach their max value very quickly (see Figure 1), as in our case they only needed eight out of the 20 possible ties to find a configuration that had the maximum number of either of the statistics. This implies that if we observe a homophily/attribute-receiver count equal to 6, 7, or 8, there could be any number of ties between 6 to 18, 7 to 19, or 8 to 20, respectively.

**Collinearity**. Following Duxbury (2021), I calculated the VIF and the correlation between transitive triads and mutual ties for networks of size five across various combinations of model parameters. The results are shown in Figure 3. Like in the previous figure, instead of simulating networks, we use the full support of the model. Furthermore, the networks used by Krivitsky et al. included between two to eight individuals, which is a major difference from Duxbury's simulation study, which featured networks with 75 nodes.

From the figure, it is apparent that, as suggested by Duxbury, the VIF rule of thumb using VIF=10 to dictate collinearity is quickly reached in ERGMs. In my example, as soon as $\boldsymbol{\theta}_{\text{mutual}}$ is above 0.5, the VIF is already above 10. Moreover, in a model with $\boldsymbol{\theta}_{\text{ttriad}} = 0$ and $\boldsymbol{\theta}_{\text{mutual}} = 1$, the VIF is about 4,600, and the correlation between the two network statistics reaches almost 1.0; a severe level of collinearity between these two terms. Notably, since in Krivitsky et al. (2022) the average network density was 0.93 and 0.73 for the $H$ (*household*) and $E$ (*egocentric*) respectively, it is to expect to observe severe VIF and correlation as that observed here.

## 4 Discussion

Krivitsky, Coletti & Hens's work is an important contribution to multi-ERG models. Their paper provides a start-to-finish framework for fitting ERGMs featuring heterogeneous sam-

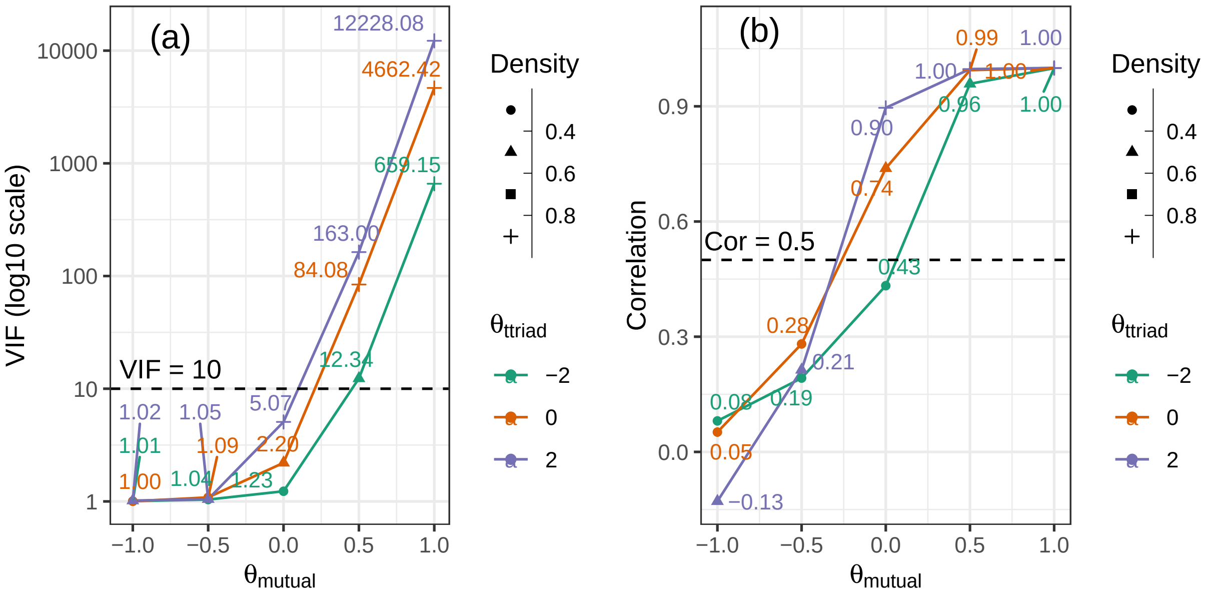


Figure 3: **VIF and correlation between terms for networks of size 5**. Each point represents the value of the VIF/Correlation for the given combination of model parameters. The VIF/Correlation is calculated using the full support of the model; in other words, instead of using simulated values as in Duxbury (2021), we use the full support of the model. The reported network density is the average density of the full support of the model.

ples of networks. Notwithstanding we got a great headstart for this type of model, there is still much work to do, especially in the case of small networks. In this paper, I have discussed two missing pieces from their work: power analysis and collinearity.

Power analysis is a missing piece in the ERGM literature. While sample size has been discussed from the inferential point of view, not much has been said about the number of networks needed to detect a given effect size. With more and more data sources featuring multiple networks as in Krivitsky et al. (2022), it is important to have a way of justifying the number of networks needed to detect a given effect size. This is especially critical from the research funding point of view, as such information can be critical to justifying the financial resources required for multi-network studies. Here, I propose simulation studies as a way to conduct power analyses.

In the case of collinearity, I have shown that the VIFs and correlation between terms are very high, even for small networks. Although VIFs and correlation greatly matter across linear models, after this experiment, Duxbury's work, and Krivitsky, Coletti & Hens, it is not conclusive how much it matters for ERGMs. After all, ERGMs are, by construction, endogenous.

# 5 References

# References

Butts, C. T. (2009), 'Revisiting the Foundations of Network Analysis', *Science* **325**(5939), 414–416.

Duxbury, S. W. (2021), 'Diagnosing Multicollinearity in Exponential Random Graph Models', *Sociological Methods & Research* **50**(2), 491–530.

Duxbury, S. W. & Wertsching, J. (2023), 'Scaling bias in pooled exponential random graph models', *Social Networks* **74**, 19–30.

Frank, O. & Strauss, D. (1986), 'Markov graphs', *Journal of the American Statistical Association* **81**(395), 832–842.
**URL:** *http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1986.10478342*

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N. & Morris, M. (2023), *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*, The Statnet Project (`https://statnet.org`). R package version 4.5.0.
**URL:** *https://CRAN.R-project.org/package=ergm*

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. & Morris, M. (2008), 'ergm: A package to fit, simulate and diagnose exponential-family models for networks', *Journal of Statistical Software* **24**(3), 1–29.

Krivitsky, P. N., Coletti, P. & Hens, N. (2022), A Tale of Two Datasets: Representativeness and Generalisability of Inference for Samples of Networks, Technical report.

Krivitsky, P. N., Hunter, D. R., Morris, M. & Klumb, C. (2023), 'ergm 4: New features for analyzing exponential-family random graph models', *Journal of Statistical Software* **105**(6), 1–44.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *https://www.R-project.org/*

Schweinberger, M., Krivitsky, P. N. & Butts, C. T. (2017), 'A note on the role of projectivity in likelihood-based inference for random graph models', pp. 1–6.

Schweinberger, M., Krivitsky, P. N., Butts, C. T. & Stewart, J. R. (2020), 'Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios', *Statistical Science* **35**(4), 627–662.

Vega Yon, G. (2023), *ergmito: Exponential Random Graph Models for Small Networks.* R package version 0.3-1.

**URL:** *https://cran.r-project.org/package=ergmito*

Vega Yon, G. G., Slaughter, A. & de la Haye, K. (2021), 'Exponential random graph models for little networks', *Social Networks* **64**(August 2020), 225–238.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

**URL:** *https://ggplot2.tidyverse.org*