A comment on "Tale of Two Datasets: Representativeness and Generalisability of Inference for Samples of Networks" by Krivitsky, Coletti & Hens

The recent work by Krivitsky, Coletti & Hens [KCH] provides an important new contribution to the Exponential-Family Random Graph Models [ERGMs], 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 and multi-level ERGMs, see Duxbury & Wertsching (2023), Wang et al. (2013), Slaughter & Koehly (2016)), 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-ERGMs with heterogeneous data sources. In this comment, I will focus on two issues the authors did not discuss, namely, sample size requirements and multicollinearity.

1 Sample size requirements in ERGMs

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 financial point of view, small networks are easier to collect. In ERGMs, power analysis has two different 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 ERGM, has two sides: network boundaries (as "finite" versus "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 many ERGMs are not projective, it has been shown that projectivity is not a necessary condition for consistency in likelihood-based inference (Schweinberger et al. 2017, 2020). In the case of KCH, since they deal with networks within households, boundaries are well defined, so the relevant size question is "How many households to survey?".

Number of networks. Here, 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 given the l-th statistic, $\boldsymbol{g}(\boldsymbol{y})_l = s_l$, can be calculated as follows:

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

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 $\kappa_{\mathcal{Y}}\left(\boldsymbol{\theta}\right)_{-l}=\sum_{\boldsymbol{y}'\in\mathcal{Y}:\boldsymbol{g}(\boldsymbol{y}')_l=s_l}\exp\left\{\boldsymbol{\theta}_{-l}{}^{\boldsymbol{t}}\boldsymbol{g}\left(\boldsymbol{y}'\right)_{-l}\right\}$ is the normalizing constant. In other words, once we condition on the l-th sufficient statistic, the probability mass function becomes invariant to the value of $\boldsymbol{\theta}_l$. Using the conditional distribution, we could simulate networks with a given edge count (which drives an important part of the ERG probability) and infer the required sample size. For instance, in a study where gender homophily in networks of size 8 is the main focus, and assuming an effect size of $\boldsymbol{\theta}_{\text{homophily}}=2$ and edge count of 20 (*i.e.*, a density of $(2\times 20)/(8\times 7)\approx 0.71$), we could approximate the required sample size in the following fashion:

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

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

   (b) For each set, fit a conditional ERGM to estimate $\hat{\boldsymbol{\theta}}_{\text{homophily}}$, 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 model to estimate the sample size as a function of the power, *i.e.*, $n\sim f(p_n)$.

3. With the previous model in hand, we can estimate the sample size required to detect a given effect size with a given power.

Again, it is worth emphasizing that, by fixing the number of observed ties (edge count,) we do not need to assume a value for its corresponding parameter; greatly simplifying the analysis. 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 explaining the financial resources required for multi-network studies.

2 Collinearity in small networks

Since most of the statistics used in ERGMs 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 in 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*, $\sum_{i \neq j} y_{ij} \mathbf{1}(x_i = x_j)$; (d) and *Attribute-Receiver effect*, $\sum_{i \neq j} y_{ij} x_j$, with the latter two using the binary nodal attribute *Gender*, x_i, which equals to one if the i-th node is female and zero otherwise. 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).

In-sample, not out-of-sample predictive power. Figure 2 depicts the 95% Confidence

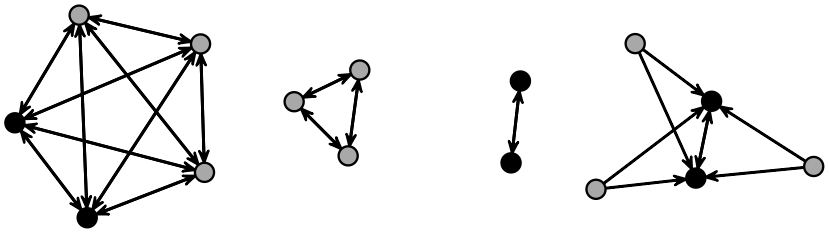

Figure 1: **Networks with maximum observed counts.** All three networks have five vertices composed of two females (black vertices) and three males (gray vertices). Network (a) shows the required set of ties to reach the maximum value of *Mutuals* and *Transitive Triads*, which coincides with a fully 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.

Interval [CI] (shaded area), and fiftieth-percentile (red dashed-lines), of the number of transitive triads (y-axis) when conditioning on the other statistics (x-axis), while at the same time, either assuming that the number of transitive triads has no effect in the data-generating-process [DGP], ($\boldsymbol{\theta}_{\text{ttriad}} = 0$), first-row of figures, or on the contrary, assuming that the term is part of the DGP, (fixing $\boldsymbol{\theta}_{\text{ttriad}} = 1$), second-row of 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

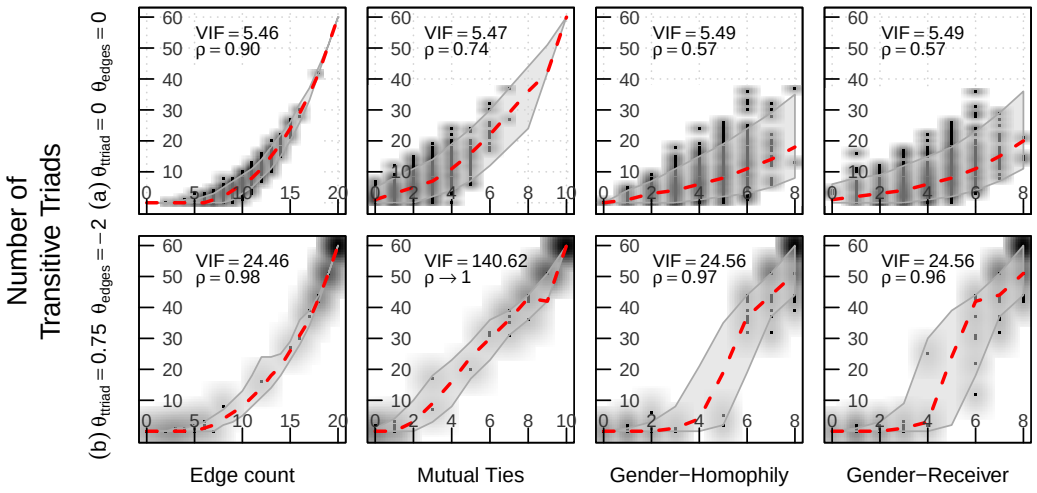

Figure 2: Conditional distribution of the number of transitive triads sufficient statistic. Row **(a)** shows the distribution under the Bernoulli model; this is, the parameter of transitivity is set to zero, whereas row **(b)** shows the distribution of the statistic assuming $\boldsymbol{\theta}_{ttriad} = 0.75$ and $\boldsymbol{\theta}_{edges} = -2$.

highly constrained. Neither *Gender-Homophily* nor *Gender-Receiver Effect* seem to be very related to transitivity; yet, as $\boldsymbol{\theta}_{\text{ttriad}} : 0 \to 0.75$, 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 140 and the correlation between those two is over 0.90. Even with this simple model, a VIF of 140 could be problematic since, as suggested by Duxbury (2021), VIFs above 150 should be concerning.

Concerning the relatively poor predictive power of the *Gender-Homophily* and *Gender-Receiver Effect* sufficient statistics, part of this could be because they reach their maximum 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.

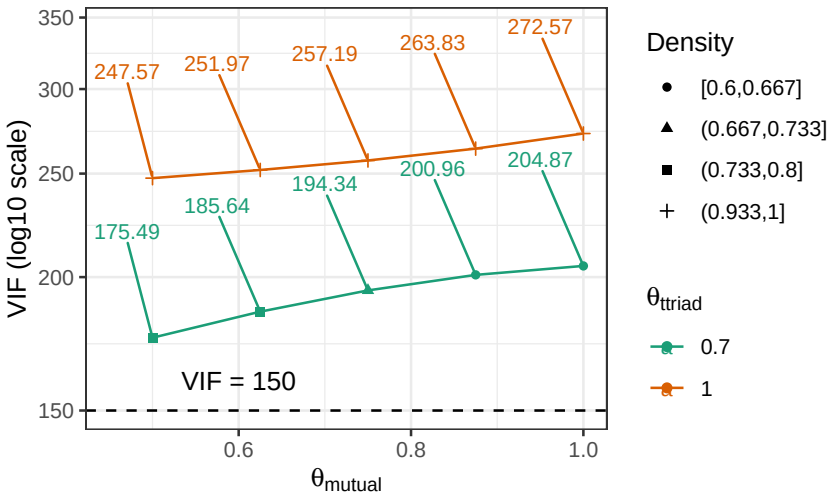

Figure 3: **VIF for transitive triads for networks of size 5**. Each point represents the value of the VIF for the given combination of model parameters. The VIF is calculated using the full support of the model; in other words, instead of using simulated values as in Duxbury (2021), we use the full support of the model. The reported network density is the average density of the full support of the model. The models include an edge parameter raging between -2 and -3 to control for network density.

Collinearity. Following Duxbury (2021), I calculated the VIF and the correlation between transitive triads and mutual ties for networks of size five (with KCH networks' between 2 and 8 nodes) 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.

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 this example, all VIFs are above the 150 threshold suggested by Duxbury, and correlation between transitivity and mutuality close to one; including in networks with densities below 0.66. 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.

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.

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.

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.

Slaughter, A. J. & Koehly, L. M. (2016), 'Multilevel models for social networks: Hierarchical Bayesian approaches to exponential random graph modeling', *Social Networks* **44**, 334–345.

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.

Wang, P., Robins, G., Pattison, P. & Lazega, E. (2013), 'Exponential random graph models for multilevel networks', *Social Networks* **35**(1), 96–115.

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

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