Diagnosing Multicollinearity in Exponential Random Graph Models

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of The Ohio State University

By

Scott W Duxbury

Graduate Program in Sociology

The Ohio State University

2017

Master's Examination Committee:

Dana L Haynie, Advisor

David Melamed, Co-Advisor

Christopher Browning

Abstract

Exponential random graph models (ERGM) have been widely applied in the social sciences in the past ten years. However, diagnostics for ERGM have lagged behind their use. Collinearity-type problems can emerge without detection when fitting ERGM, skewing coefficients, biasing standard errors, and yielding inconsistent model estimates. This leads to a unique paradox in statistical models of social networks: as more endogenous network effects are modeled, the likelihood of encountering poor model estimates may also increase. This paper provides a method to detect multicollinearity when using ERGM. It outlines the problem and provides a method to estimate shared variance between ERGM parameters. It then tests the method with a Monte Carlo simulation, fitting 27,000 ERGMs and calculating the variance inflation factors for each model. The distribution of variance inflation factors is analyzed using multilevel regression to determine what network characteristics lend themselves to collinearity-type problems. The parameter space of these variables is then examined to specify at what variance inflation factor value a researcher may expect problematic multicollinearity.

Acknowledgments

I would like to thank David Melamed, Dana Haynie, David Schaefer, Christopher Browning, Jacob Young, and Eric Schoon for advising, critique, compliments, and encouragement. I would also like to thank Jenna Johnson for support.

Vita

May 2009 .....East Lansing High School

2015 .....B.A. Sociology, Anthropology, Western Michigan University

2015 to present .....Graduate Student, Department of Sociology, The Ohio State University

Publications

Duxbury, Scott W. (Forthcoming). "Information Creation on Online Drug Forums: How drug use becomes moral on the margins of science." *Current Sociology*. DOI: 10.1177/0011392115596055.

Fields of Study

Major Field: Sociology

Table of Contents

List of Tables

List of Figures

Section 1: Introduction

Advances in the statistical modeling of social networks have provided a much needed analytic tool for social scientists. The exponential random graph model (ERGM) has been widely applied to test theories of racial boundaries (Wimmer and Lewis 2010), group processes (Goodreau, Kitts, and Morris 2009), the formation of drug distribution networks (Duxbury and Haynie unpublished manuscript), gang violence (Papachristos 2009; Papachristos, Hureau, and Braga 2013), and social structure in total institutions (Kreager et al., unpublished manuscript), among other topics. The main benefit of ERGM is the ability to simultaneously model and compare actor level attributes against the effects of endogenous network substructures (Robins et al. 2007; Lusher, Robins and Koskinen 2013).[1] While the theory behind these models has been well established for some time (Holland and Leinhardt 1981; Frank and Strauss 1986), researchers have only begun to widely use ERGM in the past 10 years. This is largely because of diagnostic breakthroughs that allow researchers to evaluate model fit (Handcock 2003; Hunter et al. 2003; Hunter, Goodreau, and Handcock 2008) and Bayesian estimation methods to derive independent parameter estimates from network data (Snijders 2002; Handcock 2003).

[1] In ERGM, an endogenous effect is an explanatory variable related to the social structure of the network.

Despite widespread use of ERGM, the variety of diagnostic tools commonly used to evaluate bias in more conventional statistical models have yet to be adapted for ERGM. Particularly, the issue of multicollinearity often goes unchecked and, by extension, undetected. The lack of multicollinearity diagnostics for statistical network models is not surprising. The standard approach to diagnose multicollinearity in a regression model—the variance inflation factor (VIF)—implies independent observations. Consequently, translating this tool to network data is an unclear task. Moreover, extremely high multicollinearity in an ERGM may lead to model degeneracy and non-convergence (Lusher, Koskinen, and Roberts 2013; Chandrasekhar and Jackson 2014).

These practical limitations do not subvert the issue that multicollinearity in ERGM can yield similar estimation problems as in any statistical model: skewed regression coefficients and biased standard errors. Moreover, unique collinearity-type problems arise when using ERGM (Snijders et al. 2006; Dekker, Krakhardt, and Snijders 2007; Salter-Townshend and Murphy 2015). First, the reliance of dyadic dependent ERGMs on Markov Chain Monte Carlo (MCMC) maximum likelihood estimation (MLE) means that highly collinear estimates may result in multiple models from which the best fitted model cannot be determined (Chandrasekhar and Jackson 2014). Similarly, because endogenous network parameters all often vary in tandem with the networks' density, an increasing number of structural terms may increase the likelihood of collinearity-type problems (e.g. Hunter 2007). This is particularly problematic since the inclusion of endogenous parameters is often the primary appeal of using ERGM.

How can the VIF be translated to ERGM? What VIF value would indicate problematic levels of multicollinearity? What ERGM specifications are vulnerable to collinearity-type problems? This paper proposes a method to diagnose multicollinearity in ERGM. It introduces a technique to robustly estimate the shared variance between variables using a distribution of networks simulated from the parameters of a converged ERGM. It then provides the results from a Monte Carlo simulation experiment designed to create a distribution of VIFs retrieved from ERGM. The method is tested in the simulation experiment and the ERGM specifications which correlate with large VIFs are identified using multilevel regression. Based on the results of the simulations, a new rule of thumb for evaluating problematic multicollinearity in ERGM is provided: a VIF>110 is concerning, a VIF>1100 is problematic.

Section 2: Problems Diagnosing Multicollinearity in Exponential Random Graph Models

Formally, ERGM is defined as:

(1)

$$\Pr(\mathbf{Y} = \boldsymbol{y}) = \left(\frac{1}{k}\right) \exp\{\sum \eta_A \, g_A(\mathbf{y})\}$$

Where *A* is a type of network configuration, η_A is the parameter term, $g_A(\mathbf{y})$ is the network statistic, and *k* is a normalizing constant that ensures interpretable results and that the probability model sums to 1. Coefficients represent the log-odds of an edge connecting two vertices and decreasing Bayesian and Akaike information criteria indicate better model fit.

While standard logistic regression can predict the log-odds of binary outcomes, the dependency of network observations skews standard errors when network ties are treated as the dependent variable. Maximum pseudo-likelihood estimation has long been a tool to generate robust standard error estimates when dyads can be treated as independent from one another (Frank and Strauss 1986). However, this assumption is violated in many social networks where network substructures influence the odds of tie formation (Geyer and Thompson 1992; Robins et al. 2007; Lusher et al. 2013). This problem led to the development of MCMC MLE to generate standard errors in ERGM (Snijders 2002; Handcock et al. 2003).

The general approach to fitting an ERGM is to assume that the empirical network is the result of a stochastic process. A probability distribution of networks is then simulated by starting with the statistics of maximum-pseudo likelihood estimates using MCMC simulation to update the parameter vector. As simulated networks converge on the parameterized sufficient statistics in the network, the parameter vector converges, yielding interpretable estimates. The log-likelihood of the ERGM is evaluated iteratively until it reaches convergence. The MCMC procedure is often repeated many times to ensure that the solution to the log-likelihood is a global maximum rather than a local one. The benefit of this approach is that dyadic dependent processes, such as triadic closure or the influence of shared partnerships, can be accurately estimated.

Multicollinearity may be particularly difficult to detect with the inclusion of dyadic dependent terms (Snijders et al. 2004; Salter-Townshend and Murphy 2015). This is because endogenous network statistics largely depend on the network's density, which functions equivalently to the intercept in OLS regression (Robins et al. 2007; Faust 2007; Lusher et al. 2013). Consequently, the correlation between endogenous and exogenous predictors may be inflated by an endogenous term's inclusion or in higher density networks. Compounding this problem, multicollinearity may dramatically increase with the inclusion of multiple endogenous terms, all of which depend on the networks' density, and in which higher order substructures depend on lower-order ones (Faust 2007).

Calculating shared variance between parameters

One large problem in detecting multicollinearity in ERGM relates to translating VIFs to the ERGM context. VIFs in GLM are based on the assumption of accurate and otherwise independent correlations between predictor variables (Fox and Monelli 1992; Fox 2008). The general approach is to regress the term of interest on all other explanatory variables in the model using OLS. The resulting R^2 to is used to calculate a VIF for that variable. In this approach, R^2 is calculated as:

(2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Where *y* is the outcome variable and $\hat{y}$ is the predicted value of *y*. VIF is calculated as:

(3)

$$VIF = \frac{1}{1 - R^2}$$

This approach works well in GLMs because observations are independent and errors are assumed to be uncorrelated. When predictor variables are highly correlated, variance estimates inflate and can be subsequently detected. However, variance estimates resulting from an OLS regression are inaccurate when working with network data because observations are not *iid*. Utilizing this method on network data may lead to R^2 values that do not reflect shared variance among network parameters, and therefore VIFs that do not accurately estimate multidimensional correlations between explanatory variables. This problem is likely to increase as dependencies between dyads are included, and so the method will be especially inaccurate in ERGMs with endogenous parameters.

The proposed alternative for ERGMs is to simply calculate R^2 from the fitted parameters of an existing model. This approach is particularly appealing for statistical network models, which explicitly seek to evaluate independent correlations between variables when using non-*iid* data. By calculating R^2 directly from the correlation matrix, resulting VIFs will capture shared variance between explanatory variables.

This method assumes an ERGM has already converged. After yielding estimates, a distribution of networks can be simulated from the parameters of the ERGM using MCMC techniques (Geyer and Thompson 1992; Hunter et al. 2008). An accurate bivariate correlation matrix $\boldsymbol{R}$ can be calculated from the statistics of these networks. $\boldsymbol{R}$ can then be broken into a vector of correlations $\boldsymbol{r_{xy}}$, where subscript *y* represents the explanatory variable of interest and *x* represents all other explanatory variables in the ERGM, and a square correlation matrix $\boldsymbol{R_{xx}}$ consisting of all explanatory variables in the model other than *y*. Correlations with the edges term should be excluded from both $\boldsymbol{r_{xy}}$ and $\boldsymbol{R_{xx}}$. Thus, in the fitted ERGM, R^2 for an explanatory variable of interest *y* be calculated as (Rencher 2002):

(4)

$$R^2 = \boldsymbol{r_{xy}^T R_{xx}^{-1} r_{xy}}^2$$

This R^2 can then be used in Eq. (3) to calculate the VIF for the ERGM.

It may seem curious to simulate a distribution of networks rather than to calculate $\boldsymbol{R}$ from the ERGM directly. MCMC MLE explores unlikely network configurations when

[2] This can be reduced to Eq. (2) for a more parsimonious equation in the regression context. However, this reduction changes the input values on the right-hand side of the equation from correlations reflecting independent parameters to values observed in independent observations.

approaching convergence. Consequently, the correlation matrix of a fit ERGM may reflect some error implicit in the simulation procedure. Thus, calculating R^2 directly from the prior distribution of ERGM networks may yield biased estimates. It is therefore important that the researcher simulate a distribution of networks from a converged ERGM and use this posterior distribution of networks to calculate $\boldsymbol{R}$.

To summarize, the proposed approach is to:

1) Fit an ERGM.
2) Simulate a large distribution of networks from the ERGM parameters.[3]
3) Calculate R^2 using Eq. (4), treating the explanatory variable of interest as *y*.
4) Calculate VIF using Eq. (3).

To evaluate this method, a single condition Monte Carlo simulation experiment that regressed a randomized variable on three randomized variables using OLS regression was compared to a dyadic independent ERGM.[4] The ERGM was Bernoulli, predicting a 75-vertex network with three randomized node-level variables. Explanatory variables in both models were equidispersed (ratio of variance to mean is equal to 1). In the OLS regression, VIF was calculated using Eq. (2) and Eq. (3); the ERGM used the approach described above. The distribution of VIFs for the OLS regression yielded a mean of 2.71, a standard deviation of 0.44, and a range of 1.80 to 4.41. The distribution of VIFs for ERGM yielded a mean of 2.74, a standard deviation of 0.41, and a range of 1.81 to 4.19.

[3] A distribution of 1,000 networks is sufficient in most cases, but the number of networks should be increased in larger networks or ones with potentially problematic coefficients.
[4] Both models were ran 100 times to calculate three VIFs for 100 different OLS regressions and ERGMs. The ERGM correlation matrix was calculated from a posterior distribution of 1000 simulated networks.

The two distributions were positively correlated with a correlation coefficient of 0.13. These results indicate that the proposed method is comparable to well-known methods and yields similar results even when applied to non-independent outcomes.

It is worth discussing here the meaning of this VIF in the ERGM context. There currently is not a clear estimate of variance explained for ERGM parameters, including the R^2 in Eq. (4). 'Spectral' goodness of fit measures for social networks have only been developed recently (Shore and Lubin 2015) and they do not easily translate to estimating multicollinearity. Instead, the calculation for R^2 proposed in Eq. (4) estimates shared variance; where increasing values of the ERGM VIFs indicate higher multivariate correlations between explanatory variables. Thus, while increasing VIFs across models indicate increasing multidimensional correlations, it is not clear how VIF values should be interpreted on their own.

In order to evaluate how these VIF values fluctuate—i.e. when a VIF is high enough to warrant concern—the properties of ERGM VIFs must be examined. The remainder of this paper is dedicated to 1) defining when VIF values are problematic and 2) identifying ERGM specifications that are vulnerable to multicollinearity.

Evaluating ERGM VIFs

Having defined a robust method to calculate the VIF from network data, ERGM specific problems resulting from dyadic dependencies can be directly addressed. The main ambiguity in evaluating VIFs in ERGM is that MCMC estimation for ERGM is much more tolerant to correlation between explanatory variables. MCMC estimation

expects a relatively high degree of collinearity between endogenous parameters (Snijders 2002; Robins et al. 2007).[5] Consequently, the standard rule of thumb for evaluating VIFs—where a VIF larger than 4 indicates concerning levels of collinearity, and a VIF greater than 10 indicates problematic levels of collinearity (Chatterjee and Price 1991; O'Brien 2007)—does not translate well to the ERGM context because VIFs in ERGM will likely be much larger. Similarly, the VIF in ERGM does not have a straightforward interpretation as in GLM, where $\sqrt{VIF}$ is the factor by which variance estimates are inflated.

How should researchers interpret VIFs in ERGM? What ERGM specifications may increase the risk of multicollinearity? To answer these questions, I conducted a 27 factorial condition Monte Carlo simulation experiment to investigate the properties (or distribution) of VIFs under a variety of ERGM specifications. Monte Carlo simulation generates random samples of data from fixed specifications—or factorial conditions (Mooney 1997). The underlying logic of this strategy is to create a distribution of VIFs that reflect a wide variety of ERGM specifications and network configurations. The distribution of VIFs resulting from these experiments are then analyzed using multilevel regression models to evaluate which ERGM specifications may result in high multicollinearity. Finally, results from multilevel regression are evaluated to specify at which VIF values concerning multicollinearity may be present.

[5] MCMC calculates variance using change statistics, where the values of the observed step in the chain are recorded only if the observed edge changes from 0 to 1, or vice versa, and the simulated network fulfills the acceptance criteria. Consequently, high correlations in the observed network only impact the standard errors of the ERGM if the change statistics are highly correlated as well, i.e. if the high correlation persists throughout the MCMC procedure.

Section 3: Experimental Conditions

A Monte Carlo simulation experiment with 27 conditions gives insight as to when multicollinearity skews model estimates. The simulated networks are random (e.g. Erdos and Renyi 1967) with 75 vertices. An ERGM is then fit to the network per the specifications of that factorial condition. A posterior distribution of 1,000 networks is then simulated from the fitted ERGM. A correlation matrix is retrieved from this distribution and the VIF for each variable is calculated using Eq. (4) and Eq. (3).

The factorial design of the experiment varies network density, ERGM parameters, and variance-mean ratios of node-level attributes to create a robust distribution of ERGM VIFs that reflect variation in realistic ERGM specifications (see Table 1). Each ERGM is predicted with three random node level attributes with varying variance-mean ratios. The variance-mean ratio is the ratio of variance (σ^2) to mean (μ), defined as $\frac{\sigma^2}{\mu}$. Higher variance-mean ratios lead to over dispersion, and small variance-mean ratios indicate low dispersion in the probability distribution of a model. By varying the variance-mean ratio of node-level attributes, the simulation experiment introduces (or reduces) correlation between exogenous variables. Variance-mean ratio values vary across low dispersion (0.5), equidispersion (1), and over dispersion (5). Network density varies within realistic values (0.1, 0.2, 0.3) (Steglich et al. 2010), to gain representative estimates of potential dependencies between endogenous terms (higher density will likely lead to higher levels

of collinearity in the endogenous terms). Three different endogenous terms are included: degree correlation, *gwesp,* and degree popularity. Degree correlation measures the correlation of degree scores between vertices connected by an edge; degree popularity measures an actors' degree score relative to the degree score of other actors; *gwesp* measures localized clustering through edgewise shared partnership.[6] These terms were chosen for their theoretical relevance and common usage to parse out the independent effects of global and local hierarchies in network formation (e.g. preferential attachment vs. localized clustering) (Snijders, van de Bunt, and Steglich 2010). Each condition of the experiment adds an additional endogenous term to force increasing multicollinearity between network substructures. When there is only one endogenous parameter in the model, it is *gwesp*; when there are two endogenous parameters, the model includes both degree popularity and *gwesp*; where there are three endogenous parameters, the model includes degree correlation, *gwesp,* and degree popularity.

Conditions	**Values**		
Network density	.1	.2	.3
Variance mean ratios of exogenous variables	.5	1	5
Count of endogenous parameters	1	2	3
Endogenous parameters	*gwesp* (0.7)[7]	Degree popularity	Degree correlation

Table 1. Conditions of the experiment

[6] The *gwesp* term is a geometrically weighted term for edgewise shared partnership. It falls under the 'curved' family of ERGM terms and is generally a more robust estimate than network transitivity. See Hunter (2007) for an introduction and discussion.
[7] Decay parameter of the *gwesp* term is fixed at 0.7. See Hunter (2007) for an introduction.

Section 4: Results

The distribution of ERGM VIFs

The first concern is to characterize the distribution of VIFs retrieved from the Monte Carlo experiment to understand how the proposed method behaves across a range of ERGM specifications. Unlike VIFs in GLM, ERGM estimation is much more tolerant to multicollinearity, and so models with multicollinearity problems may yield absurdly high VIFs. The range of VIFs in the experiment stretches from 1.00 to 2.75×10^7. The mean VIF value is 1077.01 with a standard deviation of 1.48×10^5. These results indicate extreme negative skew in the distribution of VIFs. Logarithmic transformation brings the VIFs into a more interpretable range. The resulting mean of log(VIF) is 1.49, with a standard deviation of 1.78 and a range of 0.00 to 17.13. These trends are demonstrated graphically in Figure 1. Over half of (log) VIF values range between 0 and 1 in value, indicating that extremely high VIF values are still relatively rare when using ERGM. Only 4.76% of log(VIF) are greater than 5. Log(VIF) is dealt with for the remainder of this paper for the sake of interpretability and consistency.

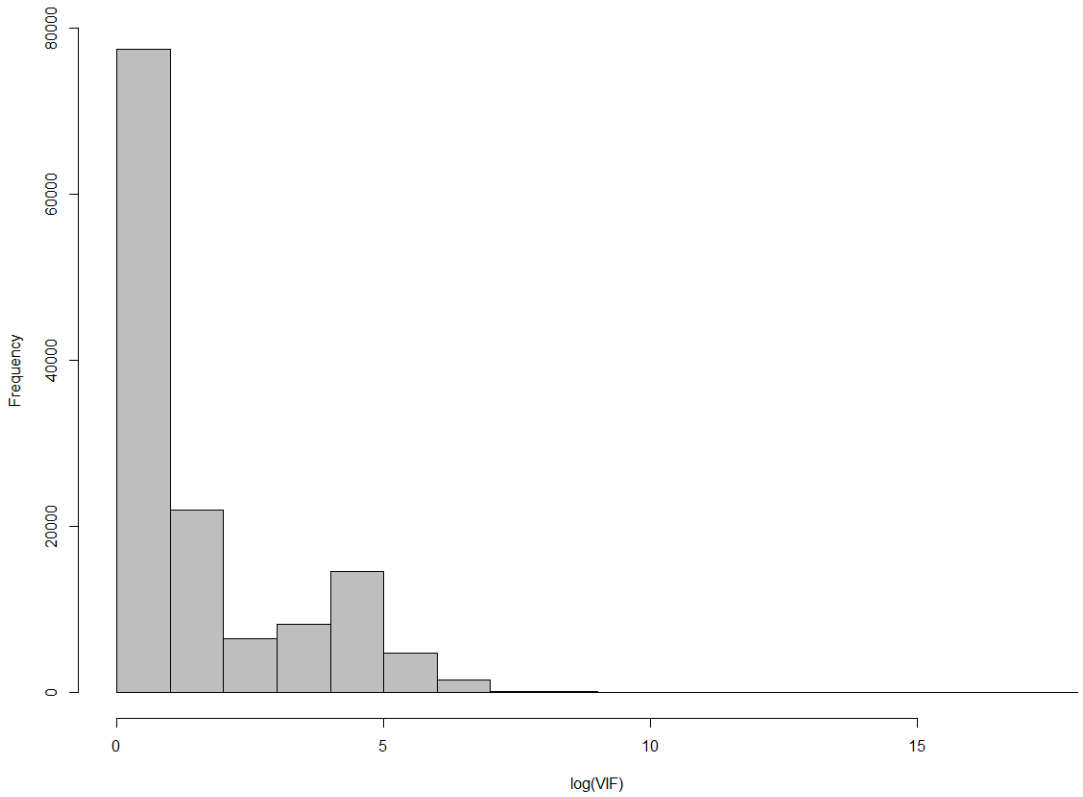

Figure 1. Distribution of log(VIF)

What ERGM specifications are likely to produce collinearity-type problems?

Prior research suggests that an increasing amount of endogenous terms may increase the likelihood of encountering collinearity-type problems (Lusher, Koskinen, and Roberts 2013; Chandrasekhar and Jackson 2014), however it has yet to be empirically examined. It is also unclear whether a simple increase in the count of endogenous parameters increases the risk of multicollinearity or if endogenous variables may spur multicollinearity more so than exogenous variables. Multicollinearity may also increase as network density increases because the specific attributes or endogenous

effects which correlate with selection patterns may be more difficult to identify as actors connect with one another more frequently. There may also be an interactive effect, where high density networks are less robust to endogenous specifications than low density networks because parameters may be highly correlated with the density of the network. These potential trends are investigated empirically using three-level linear regression on the distribution of VIFs (VIFs nested in ERGMs nested in experimental conditions). Log(VIF) is the dependent variable. Fixed effects include the type of variable, variance mean ratios, network density, the count of endogenous variables in the model; random effects include network density and the count of endogenous variables.[8] All random effects are specified at the level of conditions; the model and condition level intercepts vary randomly. While an uncommon approach, regression on simulated data is useful to evaluate multidimensional correlations that may not be apparent through descriptive statistics alone (Kim and Bearman 1997; Hanaki et al. 2007).

Model 1 includes the network density, variance mean ratios, and combinations of specific endogenous parameters. Results show that increases in density are correlated with increased VIFs. Results also indicate that compared to exogenous parameters, endogenous parameters tend to be correlated with higher VIFs. Results also indicate that as the variance-mean ratio of exogenous variables increase (towards over dispersion), log(VIF) decreases.

[8] Additional models were ran with variance mean ratios varying randomly on the third level. However, the variance components for the coefficients were <0.00001 across all models, yielding similar coefficient and standard error estimates for fixed effects.

Independent Variables	Model 1	Model 2	Model 3	Model 4
Network density	-	-	-	-
0.2	0.415 (0.163)	0.668 (0.068)	0.636 (0.066)	0.129 (0.058)
0.3	0.833 (0.238)	1.164 (0.016)	1.116 (0.009)	0.184 (0.009)
Variance mean ratios	-	-	-	-
1	-0.400 (0.155)	-0.352 (0.019)	-0.349 (0.011)	-0.349 (0.009)
5	-0.577 (0.155)	-0.487 (0.019)	-0.482 (0.012)	-0.482 (0.006)
Count of endogenous parameters	-	0.908 (0.055)	0.432 (0.050)	0.429 (0.041)
Endogenous parameters	-	-	-	-
gwesp	1.482 (0.006)	-	1.482 (0.006)	0.322 (0.009)
Degree popularity	3.169 (0.008)	-	3.169 (0.008)	1.924 (0.011)
Degree correlation	3.572 (0.010)	-	3.572 (0.010)	3.011 (0.016)
Endogenous parameters*density	-	-	-	-
0.2**gwesp*	-	-	-	1.199 (0.013)
0.3**gwesp*	-	-	-	2.281 (0.013)
0.2*degree popularity	-	-	-	1.221 (0.016)
0.3*degree popularity	-	-	-	2.513 (0.016)
0.2*degree correlation	-	-	-	0.591 (0.022)

Continued

Table 2. Multilevel regression analysis of log(VIF) in a random network.

Table 2 continued

0.3*degree correlation	-	-	-	1.092 (0.022)
Intercept	0.382 (0.116)	-0.772 (0.167)	-.718 (0.150)	-0.217* (0.122)
Variance components (std. dev.)				
Model-level intercept	0.000* (0.000)	0.000* (0.000)	0.000* (0.000)	0.000* (0.000)
Condition-level intercept	0.049* (0.221)	0.494 (0.703)	0.436 (0.816)	0.275 (0.524)
Density (0.2)	0.325* (0.569)	0.075 (0.274)	0.082 (0.285)	0.005 (0.229)
Density (0.3)	0.512* (0.716)	0.000 (0.005)	0.000 (0.001)	0.000 (0.002)
Count of endogenous parameters	-	0.055 (0.236)	0.049 (0.220)	0.031 (0.175)
AIC	3.49×10^5	4.97×10^5	3.49×10^5	3.09×10^5
R^2[9]	0.73	0.27	0.76	0.82

*Not statistically significant at the 5% level; *N* for count of VIFS is 135,000, *N* of models is 27,000, N for conditions is 27.
**Reference category for network density is 0.1; reference category for variance mean ratios is 0.5 reference category for endogenous parameters is whether the VIF observation was an exogenous variable; reference category for interaction is density (0.1)*exogenous variable. Standard errors are in parentheses.

Model 2 includes the global count of endogenous parameters and removes the factor variable for specific endogenous parameters.[10] The count of endogenous parameters is positively associated with log(VIF) for all variables in the model, indicating

[9] This R^2 is calculated using Snijders and Bosker's (2008) total model variance method.
[10] To clarify this distinction, the specific endogenous variable parameter is a nominal factor variable associated with specific values of log(VIF) at level 1; the global count of endogenous variables is a level 3 continuous variable indicating the count of endogenous variables in the ERGM.

that as more endogenous parameters are included, the overall log(VIF) for each term increases. However, the large decline in R^2 between these two models indicates that the log(VIF) values associated with endogenous parameters explain more variance in the log(VIF) outcome than the global count of endogenous parameters in the model. This is reflected in a comparison of the Akaike Information Criteria of the two models: the AIC of Model 1 is 3.49×10^5, whereas the AIC of Model 2 is 4.97×10^5.[11]

Model 3 adds both the count of endogenous parameters and the term for specific endogenous variables. The decline in strength of the count of endogenous variables indicates that the effect that an increase in endogenous specifications has on log(VIF) for each term in the model is partially explained by the increase in log(VIF) for specific endogenous variables. These results lend support to the thesis the collinearity problems are most likely to emerge among endogenous parameters, as opposed to exogenous variables.

Model 4 tests cross-level interactions between specific endogenous variables and the density factorial condition. The coefficients indicate that increasing network density and endogenous variable specifications have a positive interactive effect on log(VIF) when compared to the log(VIF) of exogenous variables in a sparse network (0.1 density). In terms of main effects, each additional endogenous variable correlates with an increase in log(VIF). When interacted with density, there is a diminishing marginal return on log(VIF) as the density increases. Figure 2 shows these trends graphically. The difference

[11] An F-test was not possible because the models are not nested. Lower AIC is commonly used as an alternative goodness of fit indicator in non-nested multilevel models (Harrell 2001).

between degree correlation and degree popularity interactions decreases as the density of the network increases; the degree popularity interaction yields higher average log(VIF) than the degree correlation interaction at high levels of network density. Similarly, as network density increases, the value of log(VIF) for the *gwesp* term increases. The interaction between network density and exogenous variables shows little variance as network density changes. These results suggest that researchers using ERGM should be particularly concerned with multicollinearity when specifying numerous endogenous effects in relatively dense networks.

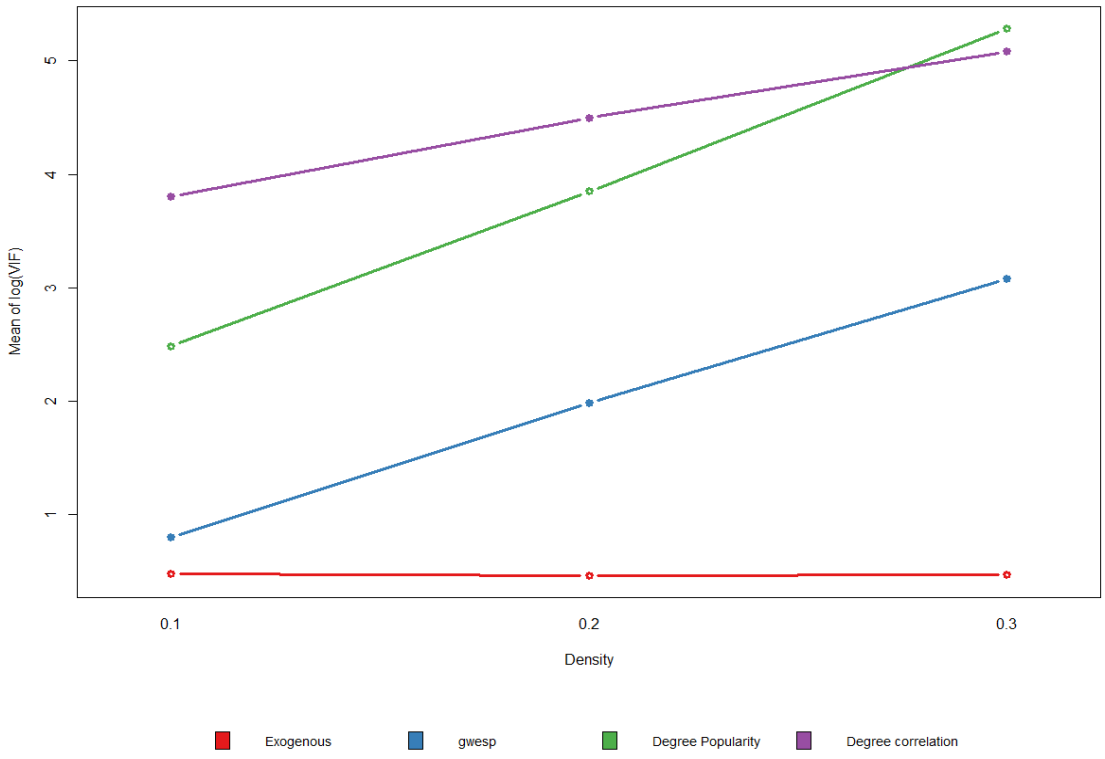

Figure 2. Interactions for variable type*density.

When is a VIF too large?

The large range of VIFs retrieved from analyses reinforce the problem proposed in the beginning of this paper: many ERGMs may converge and also yield inaccurate results. In support of this, multilevel regression shows that many common ERGM specifications may generate large VIFs when used to estimate realistic social networks. Thus, the question remains, at what VIF value should a researcher consider multicollinearity to be a problem? Having generated a representative and expansive distribution of potential VIFs, the distribution can be examined to identity those VIF values which are outliers. VIF values that are relatively rare in the distribution of VIFs and also are more likely to occur with problematic specifications reflect concerning levels of multicollinearity.

Figure 3a shows that log(VIF) for exogenous variables tend to stay well below 3.[12] Alternatively, log(VIF) for the *gwesp* parameter peaks when log(VIF)=1, when log(VIF)=3, and when log(VIF)=5. The degree popularity term peaks when log(VIF)=4.5; degree correlation specifications also show higher concentration between a log(VIF) value of 4 and 5.[13]

[12] It is important to remember that while a VIF can never be less than 1 in value, the log of VIF can be as low as 0.

[13] Both the degree popularity and degree correlation terms were included in the ERGM models with other endogenous variables. Consequently, descriptive statistics may be skewed by the influence of the count of endogenous variables.

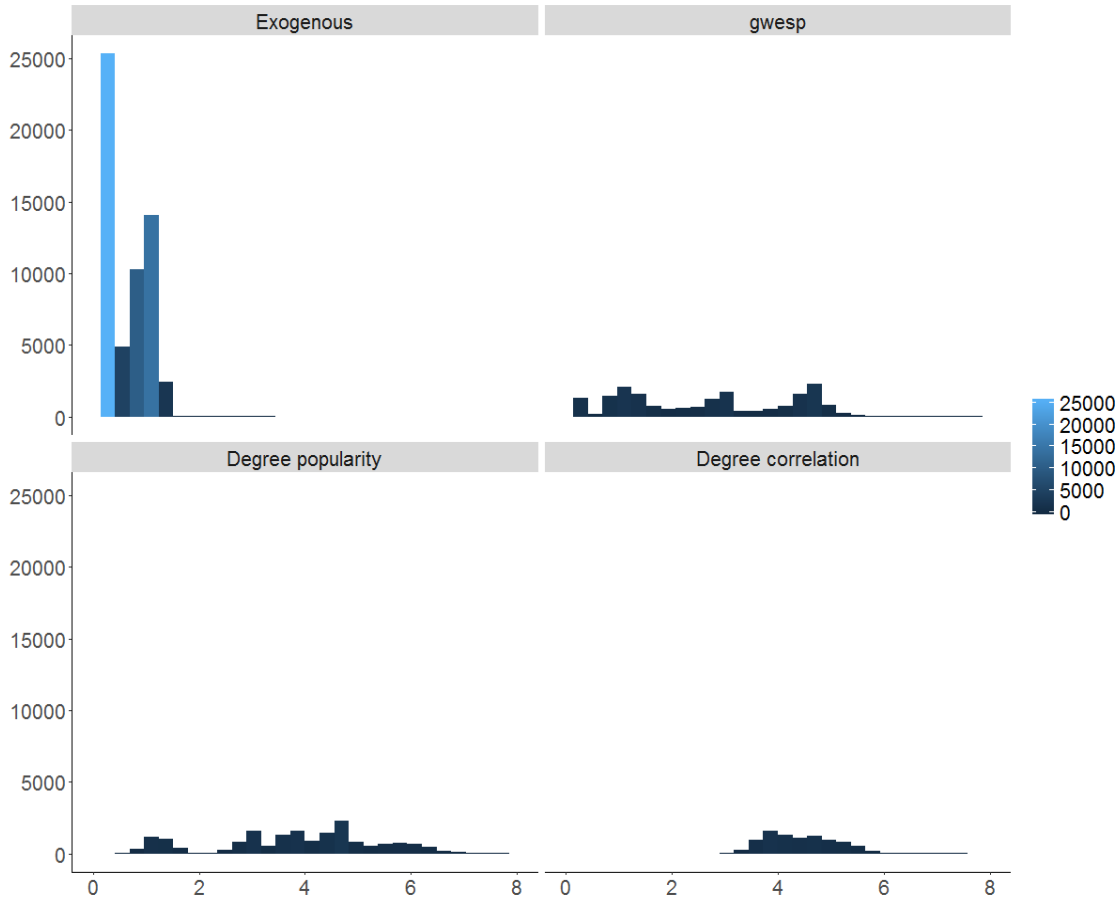

Figure 3. Frequency of log(VIF) value by variable type

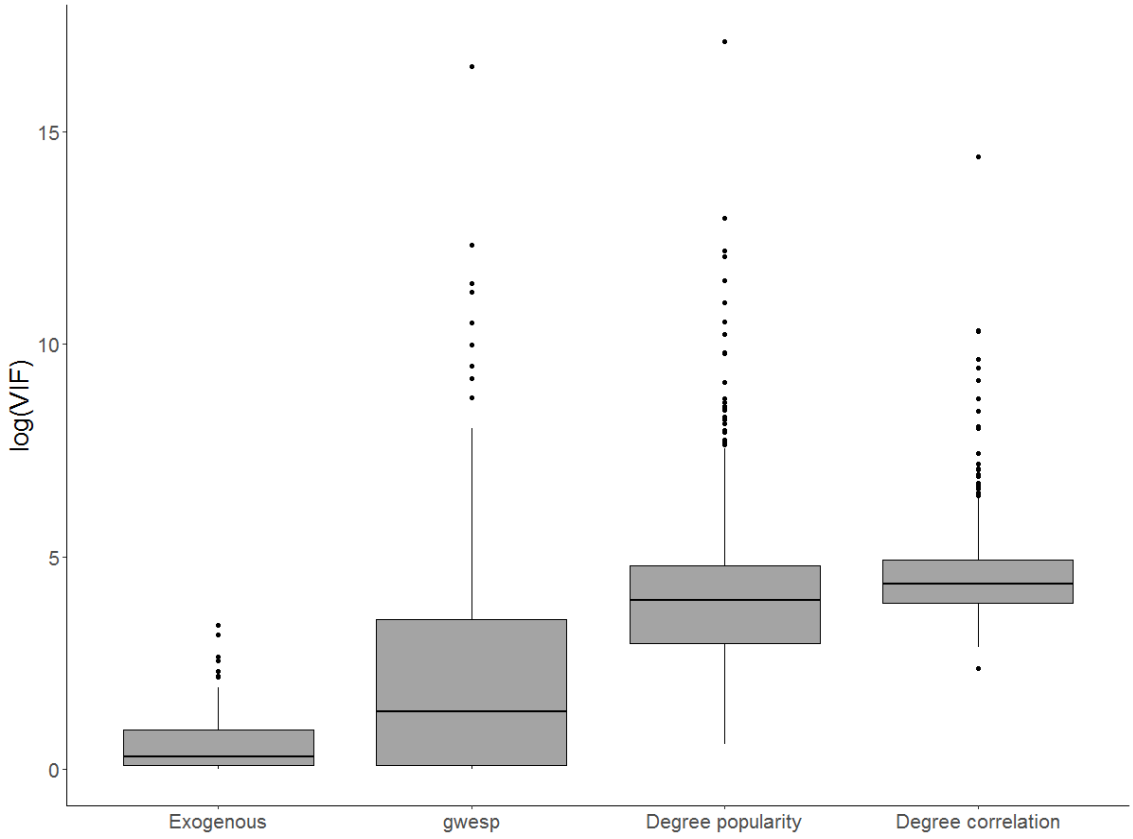

Figure 4. Frequency of log(VIF) by variable type

Log(VIF) clusters between 0 and 2 and between 4 and 5 regardless of network density (Figure 4). However, at higher values increases as network density increases, especially between a log(VIF)=4 and log(VIF)=6. This trend is also reflected in the count of endogenous variables (Figure 5). When there is only one endogenous variable in the model, log(VIF) sits comfortably below a value of 2. However, as more endogenous parameters are included in the model, log(VIF) increasingly cluster in a range of 4 – 6.

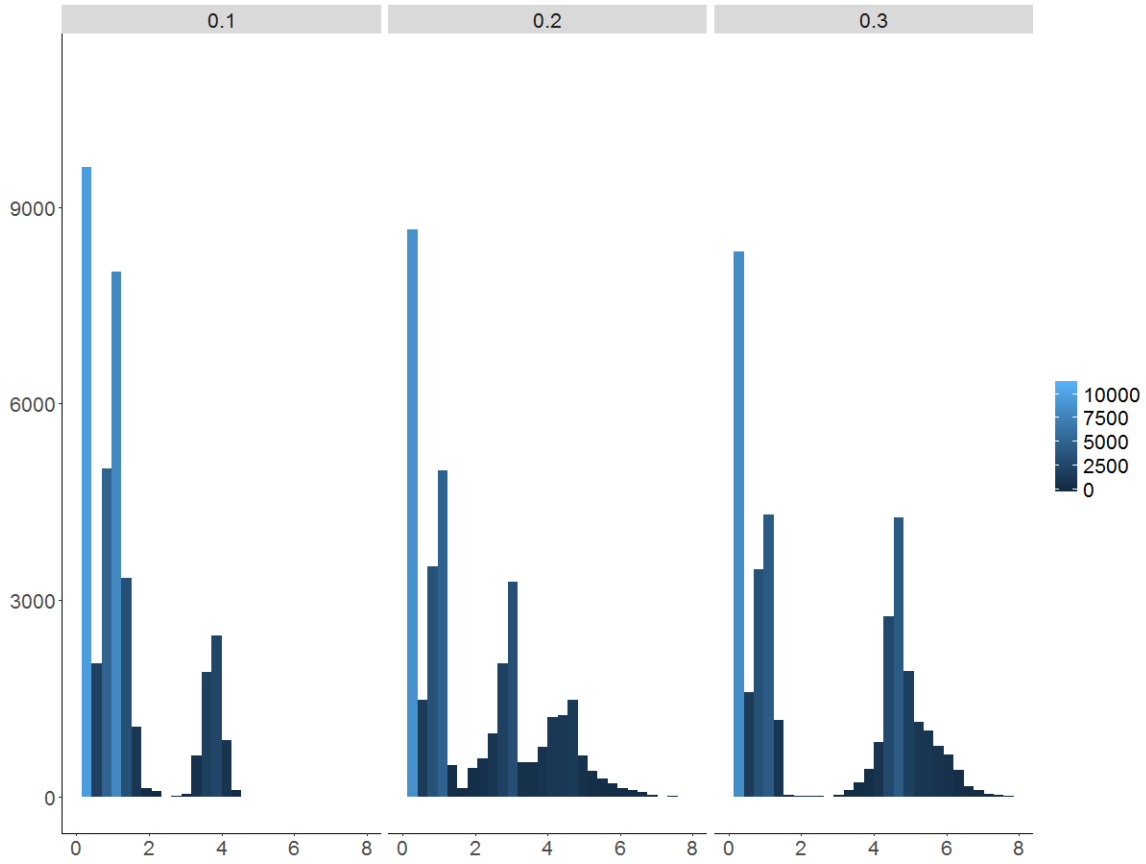

Figure 5. Frequency of log(VIF) by network density.

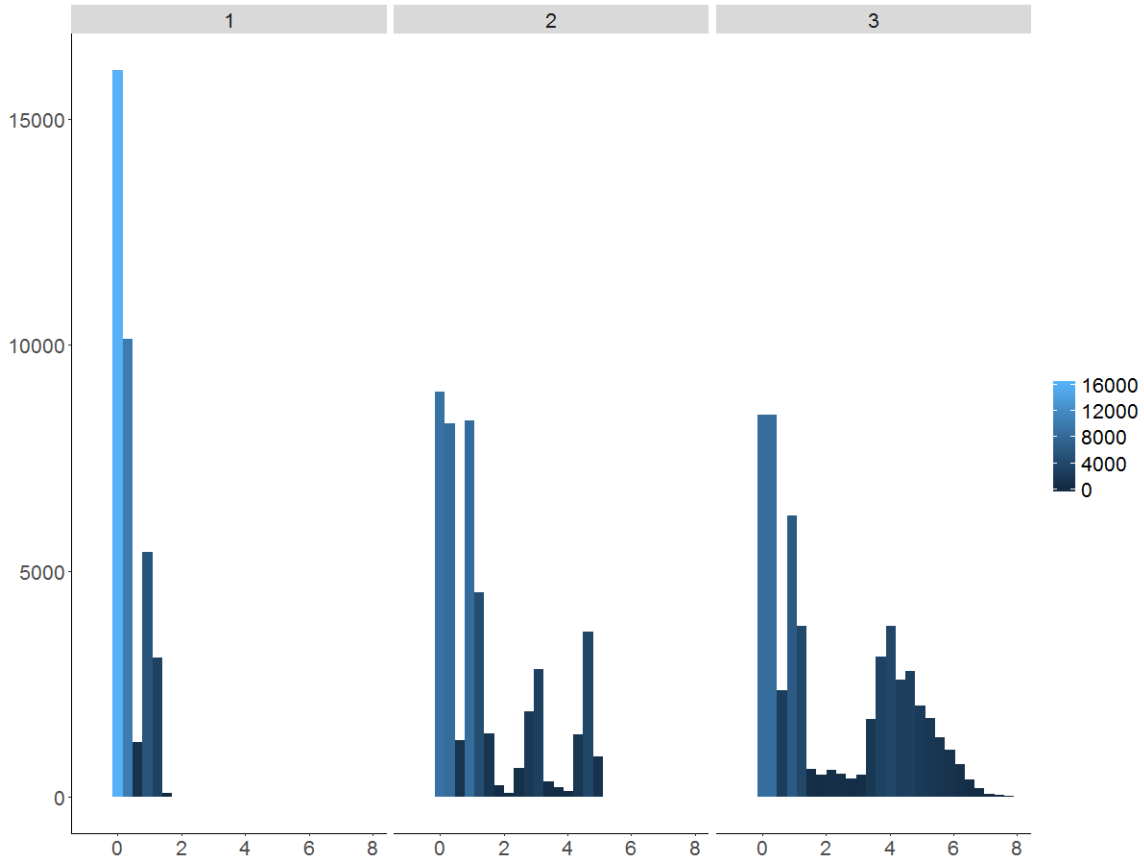

Figure 6. Frequency of log(VIF) by count of endogenous variables.

Taken in sum, these patterns reflect the predictions made in multilevel regression. They pinpoint that unproblematic network specifications (low density, low endogeneity) tend to yield VIFs in the unproblematic range, well below log(VIF)=3.[14] However, as endogeneity and network density increase, log(VIF) increasingly clusters around a value of 5. Moreover, Figure 3a demonstrates that most log(VIF) values tend to stay well below

[14] It is important to remember here that a log(VIF) of 3 still reflects high multicollinearity if seen in a GLM (exp(3)=20.086).

5 even among problematic variables (95.8% of cases). With this in mind, it is reasonable to conclude that a log(VIF) of 5 (VIF=110) in an ERGM indicates concerning levels of multicollinearity. Further, only 1.2% of observed log(VIF) were greater than 6, making these values clear outliers in the distribution of ERGM VIFs. Thus, a log(VIF) of 6 (VIF=1097) reflects problematic levels of multicollinearity and probable bias in ERGM results.

Section 5: Discussion and Future Directions

Results from a 27 condition Monte Carlo simulation experiment show considerable kurtosis in VIF values across representative range of ERGMs. VIFs were as low as 1 and as high as 27 million. The paper then sought to evaluate which ERGM specifications were responsible for high levels of multicollinearity. It identified three themes that make an ERGM more likely to encounter collinearity-type problems: high network density, models with a large count of endogenous explanatory variables, and an interactive effect in dense networks with many endogenous explanatory variables. It also identified that VIFs associated with endogenous parameters tend to be the culprit of multicollinearity much more frequently than exogenous variables.

Results from multilevel regression models were then applied to examine clustering in the distribution of VIFs. A rule of thumb was provided to evaluate when a VIF is large enough to draw concern (VIF=110; log(VIF)=5) and when a VIF is large enough to warrant model respecification (VIF=1100; log(VIF=6). Considering the remarkable over dispersion in the range of ERGM VIFs (standard deviation =1077.01; mean=1.48×10^5), it is recommended that researchers deal with the logarithmic transformation of VIFs. To be sure, this rule of thumb should not be interpreted as a stand-alone indicator of multicollinearity. Instead, this method should be used to test

whether multicollinearity may exist in the models. Since a VIF > 1100 is rare (less than 1.2% of cases), this value indicates that a VIF greater than 1100 in a real-world ERGM is an extreme outlier and should raise real concern regarding multicollinearity in the model. Similarly, a VIF > 110 indicates to a researcher that their model includes relatively highly correlated variables (present in less than 5% of simulated VIFs). Such VIF values should drive the researcher to carefully consider their theoretical models and to examine the bivariate correlation matrix of the posterior distribution of networks as well as how VIFs and standard errors change across model specifications. As with GLMs, the most common response to a high VIF will be to remove the offending variable from the ERGM.

This project offers a much-needed statistical tool to evaluate ERGM estimates. The paper demonstrates that the proposed method is robust to both network data as well as endogenous network effects. Thus, it resolves the ERGM paradox described above by allowing researchers to pinpoint which endogenous predictors exhibit high multicollinearity. Further, while the simulation experiment used to interrogate and evaluate the method was computationally intensive, the method itself requires little computational power or specialization in statistical network analysis to execute (outlined under Eq. (4)). Thus, the general approach is relatively parsimonious and should be easily applied by researchers who may not be familiar with network analysis.

Another benefit of this approach is that the general method can be translated to other cross-sectional network models, such as the quadratic assignment procedure (Krackhardt 1987, 1988), with little modification. For example, using the approach

outlined in Eq. (4), the correlation matrix of a fitted QAP regression could be used to calculate VIFs, and skew estimates provided by simulation experiments may be adjusted to evaluate multicollinearity in other statistical network models. As such, this approach takes a promising step towards developing powerful diagnostic tools for the emerging classes of models designed to handle dependent data.

The tasks facing future research are four-fold. First, future research may extend the simulations to networks of different topographies and sizes. ERGMs of networks that exhibit preferential attachment (e.g. Barabasi and Albert 1999) may be particularly vulnerable to multicollinearity in degree-related endogenous specifications because highly connected vertices may disproportionately exhibit unique traits. Similarly, networks that exhibit localized clustering (e.g. Watts and Strogatz 1998; Watts 1999) may be particularly vulnerable to multicollinearity related to substructural transitivity parameters, like *gwesp*, because desirable traits may be highly correlated with shared partnerships. Future research may include scale-free and small-world network topographies as additional factorial conditions to evaluate whether the topography of a network lends itself to collinearity-type problems. Moreover, network size may influence how endogenous parameter specifications influence multicollinearity in an ERGM, especially because ERGM generally performs poorly when estimating large networks (Thiemichen and Kauermann 2017). With greater diversity in tie formation processes, ERGMs fit to larger networks may be able to more accurately estimate the independent correlations of variables, and thus be less vulnerable to multicollinearity. In other words, a larger network may moderate the effect network density has on multicollinearity.

Second, future research can expand on these findings by determining whether the effect of the count of endogenous variables in the model is influenced depending on which endogenous parameter is included first. Particular endogenous parameters may have ripple effects that shape the size and significance of other variables in the model. Third, due to constraints on computational power and processing time, the count of endogenous terms included in the ERGMs for these simulations has been limited. The terms provided in this paper have been selected due to their widespread popularity and common usage (Handcock et al. 2003b). A promising direction for future research is to interrogate how variance estimates change with the inclusion of less common endogenous effects, such as the arc tangent or *k*-cores. These effects are sometimes applied as specific measures in research (e.g. Hipp et al. 2013), but may have less clear behaviors when included with other network terms. Similarly, selective-sorting parameters such as homophily and heterophily are widely used by networks scholars, but may have distinct behaviors.

Fourth, a final step is to adapt multicollinearity diagnostics to the burgeoning classes of dynamic network models (Snijders, van de Bunt, and Steglich 2010; Krivitsky and Handcock 2014; Hanneke, Fu, and Xing 2010). This will help extend diagnostic tools to more recent and promising developments in longitudinal network analysis. This is particularly important, as collinearity-type problems often emerge in longitudinal analyses due to autocorrelation. Multicollinearity diagnostics will help strengthen this category of models by providing a method to detect collinearity-type problems in dynamic networks due to multicollinearity. By excluding multicollinearity as a source of

bias, researchers can determine when autocorrelation may be causing collinearity-type problems—an issue that has plagued network scholars for some time (Dekker, Krackhardt, and Snijders 2007).

References

Barabasi, Albert-Laszlo and Reka Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286: 509-512.

Chandrasekhar, Arun G., and Matthew O. Jackson. 2014. “Tractable and Consistent Random Graph Models.” *Physics and Society*. DOI: arXiv:1210.7375v4

Chatterjee, Samprit and Bertram Price. 1991. Regression Analysis by Example. Second edition. New York: Wiley.

Dekker, David, David Krackhardt, and Tom A. B. Snijders. 2007. “Sensitivity of MRQAP Tests to Collinearity and Autocorrelation Conditions.” *Psychometrika* 72(4): 563-581.

Duxbury, Scott W., and Dana L. Haynie. “From the Back-Alleys to the Dark-net: New Technological Capacities for the Efficient Organization of Crime.” Unpublished Manuscript.

Faust, Katherine. 2007. “Very Local Structure in Social Networks.” *Sociological Methodology* 37(1): 209-256.

Fox, John, and Georges Monette. 1992. “Generalized Collinearity Diagnostics.” *Journal of the American Statistical Association* 87(417): 178-183.

Fox, John. 2008. *Applied Regression Analysis and Generalized Linear Models.* Second edition. Sage.

Frank, Ove, and David Strauss. 1986. “Markov graphs.” *Journal of the American Statistical Association* 81: 832-842.

Geyer, Charles J., and Elizabeth A. Thompson. 1992. “Constrained Monte Carlo Maximum Likelihood for Dependent Data.” *Journal of the Royal Statistical Society Series B* 54(3): 657-699.

Goodreau, Steven M., James A. Kitts, and Martina Morris. 2009. "Birds of a Feather, or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks." *Demography* 46(1): 103-125.

Hanaki, Nobuyuki, Alexander Peterhansl, Peter S. Dodds, Duncan J. Watts. 2007. "Cooperation in Evolving Social Networks." *Management Science* 53(7):1036-1050.

Handcock, Mark S. 2003. "Statistical models for social networks: degeneracy and inference." Pp. 229-240 in *Dynamic Social Network Modeling and Analysis,* ed. Ronald Breiger and Phillipa Pattison. National Academies Press.

Handcock, Mark S., Garry Robins, Tom Snijders, Jim Moody, and Julian Besag. 2003a. "Assessing Degeneracy in Statistical Models of Social Networks." *Journal of the American Statistical Association* 76:33-50.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2003b. ***statnet***: *Software tools for the Statistical Modeling of Network Data.* URL http://statnetproject.org

Hanneke, Steve, Wenjie Fu, and Eric P. Xing. 2010. "Discrete temporal models of social networks." *Electronic Journal of Statistics* 4:585-605.

Harrell, Franke E. 2001. *Regression Modeling Strategies.* Springer Series in Statistics.

Hipp, John R., Carter T. Butts, Ryan Acton, Nicholas N. Nagle, and Adam Boessen. 2013. "Extrapolative simulation of neighborhood networks based on population spatial distribution: Do they predict crime?" *Social Networks* 35:614-625.

Holland, Paul W., and Samuel Leinhardt. 1981. "An Exponential Family of Probability Distributions for Directed Graphs." *Journal of the American Statistical Association* 76(737): 33-50.

Hunter, David R. 2007. "Curved exponential family models for social networks." *Social Networks* 29(2): 216-230.

Hunter, David R., Steven M. Goodreau, and Mark S. Handcock. 2008. "Goodness of Fit of Social Network Models." *Journal of the American Statistical Association* 103(481): 248-258.

Hyojoung, Kim, and Peter S. Bearman. 1997. "The Structure and Dynamics of Movement Participation." *American Sociological Review* 62(1): 70-93.

Knoke, David, and Song Yang. 2008. *Social Network Analysis.* 2nd ed. Thousand Oaks, CA: Sage.

Krackhardt, David. 1987. "QAP Partialling as a test of Spuriousness." *Social Networks* 9:171- 186.

Krackhardt, David. 1988. "Predicting with Networks: Nonparametric Multiple Regression Analysis of Dyadic Data." *Social Networks* 10:359-381.

Kreager, Derek A., Jacob T. N. Young, Dana L. Haynie, Martin Bouchard, David R. Schaefer, and Gary Zajac. "Where 'Old Heads' Prevail: Inmate Hierarchy in a Men's Prison Unit." Unpublished manuscript.

Krivitsky, Pavel N., and Mark S. Handcock. 2014. "A separable model for dynamic networks." *Journal of the Royal Statistical Society B* 76(1): 29-46.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage.

Lusher, Dean, Johan Koskinen, and Garry Robins. 2013. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications.* Cambridge Press.

Moody, James. 2004. "The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999." *American Sociological Review* 69(2): 213-238.

Mooney, Christopher Z. 1997. *Monte Carlo Simulation.* Thousand Oaks, CA: Sage.

Papachristos, Andrew V. 2009. "Murder by Structure: Dominance Relations and the Social Structure of Gang Homicide." *American Journal of Sociology* 115(1): 74-128.

Papachristos, Andrew V., David Hureau, and Anthony Braga. 2013. "The Corner and the Crew: The Influence of Geography and Social Networks on Gang Violence." *American Sociological Review* 78(3): 417-447.

O'Brien, Robert M. 2007. "A Caution Regarding Rules of Thumb for Variance Inflation Factors." *Quality and Quantity* 41: 673-690.

Rencher, Alvin C. 2002. *Methods of Multivariate Analysis.* Second edition. John Wiley & Sons Inc.

Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher. 2007. An Introduction to Exponential Random Graph (*p**) Models for Social Networks. *Social Networks* 29(2): 173-191.

Salter-Townshend, Michael, and Thomas Brendan Murphy. 2015. "Role Analysis in Networks using Mixtures of Exponential Random Graph Models." *Journal of Computational Graph Statistics* 24(2): 520-538.

Shore, Jesse, and Benjamin Lubin. 2015. "Spectral goodness of fit for network models." *Social Networks* 43:16-27.

Snijders, Tom A. B., and Roel J. Bosker. 2008. *Multilevel Analysis: An introduction to basic and advanced multilevel modeling, third edition.* London: Sage.

Snijders, Tom A. B. 2002. "Markov Chain Monte Carlo estimation of exponential random graph models." *Journal of Social Structure* 3.

Snijders, Tom A. B., Phillipa E. Pattison, Garry L. Robins, Mark S. Handcock. 2006. "New specifications for exponential random graph models." *Sociological Methodology* 36(1): 99-153.

Snijders, Tom A. B., Gerhard G. van de Bunt, and Christian E.G. Steglich. 2010. "Introduction to actor-based models for network dynamics." *Social Networks* 32: 44-60.

Steglich, Christian, Tom A. B. Snijders, and Michael Pearson. 2010. "Dynamic Networks and Behavior: Separating Selection from Influence." *Sociological Methodology* 40(1): 329-393.

Thiemichen, Stephanie, and Goeran Kauermann. 2017. "Stable exponential random graph models with non-parametric components for large dense networks." *Social Networks* 49 (1): 67-80.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications.* Cambridge, UK: Cambridge University Press.

Watts, Duncan J. and Steven H. Strogatz. 1998. "Collective dynamics of 'small-world' networks." *Nature* 393: 440-442.

Wimmer, Andreas, and Kevin Lewis. 2010. "Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook." *American Journal Sociology* 116(2): 583-642.