Detecting Heterogeneity and Inferring Latent Roles in Longitudinal Networks: Supplementary Information (SI) Appendix

Benjamin W. Campbell

March 13, 2018

Contents

Overview

This is the Supplementary Information (SI) Appendix for "Detecting Heterogeneity and Inferring Latent Roles in Longitudinal Network." The goal of this appendix is to provide answers to potential questions that readers may have once completing the manuscript. This appendix contains four sections. Section 1 contains a detailed proof of concept for the ego-TERGM mirroring the simulation study of Salter-Townshend and Murphy (2015). While the extended Monte Carlo presented in the manuscript includes this proof of concept as a single iteration, those interested in additional detail may be interested in digging into this simple iteration. Section 2 contains an extended discussion of the routine for examining the role generative process. It is prudent to provide additional detail about this routine given its required assumptions and its ability to return unbiased likelihood estimates. Section 3 presents the goodness of fit (GOF) diagnostics for the pooled TERGMs presented in the manuscript. These diagnostic give the reader a sense of how well the estimated model reflects the observed network generating process. Third, and finally, Section 4 provides further intutition for the ego-TERGM in the form of pseudocode. Pseudocode is useful to provide the reader additional detail in how model estimation procedes, from start to finish.

1 Proof of Concept

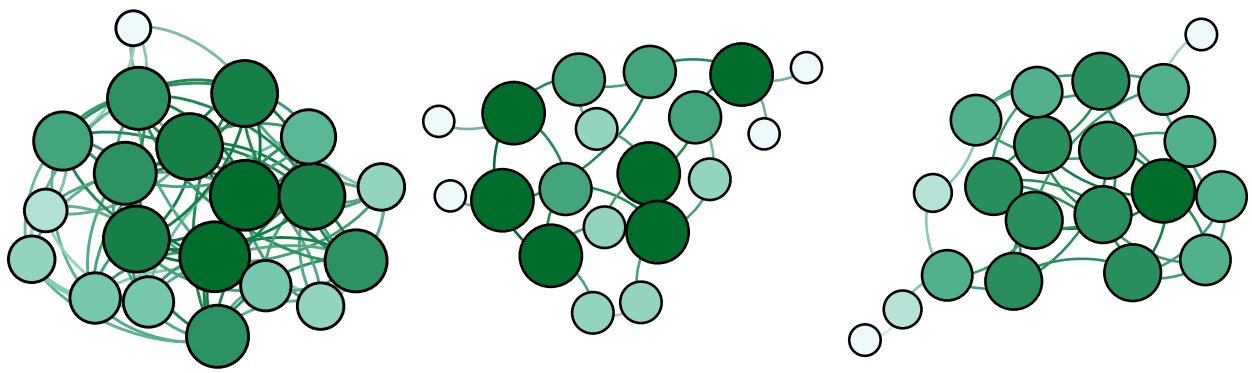
As a proof of concept, a simulation study mirroring that of Salter-Townshend and Murphy (2015) is used. In their study, they simulate 30 networks over three distinct sets of ERGM parameters, totaling 90 networks simulated according to three distinct data generating processes. My study builds upon this design by simulating five temporal observations of each of those 90 ego-networks. Following Salter-Townshend and Murphy (2015), for each of these three distinct data generating processes, a set of distinct G group-level parameters are used as specified in Table 1 where $\underline{\tau}$ refers to naive probabilities that an ego-network is assigned to a particular role defined by row, and the matrix $\boldsymbol{\theta}$ refers to simulation parameters with roles defined on the rows and parameters defined on the column in the following order: edges, geometrically weighted edgewise shared partners (GWESP) (with weight $\alpha = 0.08$) and geometrically weighted degree (GWD) (with weight $\alpha = 0.08$). These parameter values simulate networks that appear fairly similar, as demonstrated in Figure 1.

$$\underline{\tau} = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} -3 & 1 & 0 \\ -1 & -2 & -1 \\ -2 & 0 & 2 \end{bmatrix}$$

Table 1: **ERGM Parameter Values for Simulation.** Rows refer to roles and columns refer to parameters.

Overall, the model is *100% accurate in extracting correct role assignments in this simulation*, even under relatively difficult conditions. In other words, the highest probability of assignment to a particular role is always to the correct role. As Salter-Townshend and Murphy (2015) would expect, Table 2 demonstrates that group-level centroids do not reflect the simulation parameters. Once the initial ERGM fit parameters are introduced to the clustering algorithm they are transformed in such

Figure 1 Egocentric Networks Simulated According to Role Assignments. Nodes colored and sized according to degree.



(a) Simulated Role 1 **(b)** Simulated Role 2 **(c)** Simulated Role 3

a way that undermines their interpretation. As previously noted, this is the reason the ego-TERGM itself should be considered more of a clustering and role-detection model than a generative model. However, as previously discussed, pooled TERGMs may be used to return unbiased estimates of the role generating structure assuming independence across networks (an assumption met in this case). Table 3 indicates that the simulation parameters can be successfully returned using pooled TERGMs. All of the 95% bootstrap confidence intervals contain the parameters used to simulate the network. This demonstrates that one may be confident in using this routine to assess the role generative process.[1]

$$\underline{\tau} = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} 2.37 & 1.42 & 0.29 \\ -1.93 & 0.01 & 1.04 \\ -0.93 & 1.95 & -0.47 \end{bmatrix}$$

Table 2: **Estimated Parameter Values.** Rows refer to roles and columns refer to parameters.

[1]Goodness of Fit diagnostics are conducted, as recommended by Hunter, Goodreau and Handcock (2008) and Leifeld, Cranmer and Desmarais (2017), and reveal that the estimated models fit the observed data generating processes well. These diagnostics are presented later in this SI Appendix.

	Role 1 TERGM	Role 2 TERGM	Role 3 TERGM
Edges	$\mathbf{-3.11}^*$	$\mathbf{-1.03}^*$	$\mathbf{-1.98}^*$
	[−3.35; −2.91]	[−1.11; −0.96]	[−2.07; −1.89]
GWESP (0.8)	$\mathbf{1.04}^*$	$\mathbf{-1.93}^*$	$\mathbf{0.01}^*$
	[0.97; 1.12]	[−2.47; −1.71]	[−0.01; 0.03]
GW Degree (0.8)	$\mathbf{-0.47}^*$	$\mathbf{-0.93}^*$	$\mathbf{1.95}^*$
	[−1.38; 1.83]	[−1.08; −0.74]	[1.63; 2.27]
Num. obs.	65108	11863	27956

* Corresponding simulation parameter is within the 95% confidence interval

Table 3: **TERGMs fit on Pooled Group Assignments, 500 Replications**. Coefficients can be compared to the simulation parameters in Figure 1. Bolded and starred coefficients refer to estimates whose 95% confidence interval includes the simulation parameters in Figure 1.

2 Uncovering the Generative Process

As discussed in the manuscript, the ego-TERGM cannot truly be considered a generative model for roles. This is because the group-level parameter estimates, $\underline{\theta}_g$, do not resemble interpretable coefficients. In this section an approach to uncovering interpretable group-level parameters is outlined. The routine begins by estimating ego-TERGM role assignments for each ego-network. Once role assignments are extracted, longitudinal ego-networks are then pooled by common role as they are assumed to be of the same data-generating process. From that point a pooled TERGM is fit on each set of networks for $g \in G$.

While ego-TERGM cluster assignments should indicate that pooled ego-networks are identically distributed, concerns about whether they are independently distributed remain. This approach assumes that there is no temporal dependency within ego-networks or dependency across ego-networks (Cranmer and Desmarais 2011; Desmarais and Cranmer 2012). Given the use of bootstrapped MPLE, temporal dependency can be conditioned out through calculating change statistics prior to pooling (Leifeld, Cranmer and Desmarais 2017). The latter component of this assumption may not necessarily be realistic as ego-networks typically overlap (Brandes and Lerner 2007; Salter-Townshend and Murphy 2015; Box-Steffensmeier et al. 2018). *As such, when exploring this option for assessing the generative structure for each role, the independence assumptions and whether it has been met must be considered.* When both assumptions are met unbiased estimates should result.

To test whether this technique uncovers the correct data generating process we examine the simulated networks discussed in the manuscript's proof of concept. 90 networks observed over five time periods each are simulated according to three distinct data generating processes. The networks are simulated according to three sets parameters for edges, GWESP, and GWD presented Table 1. As expected the group-level centroids estimated by the ego-TERGM do not return these values.

This is confirmed by the discussion in the manuscript and the estimates presented in Table 2.[2]

When using the pooled TERGM, however, accurate estimates for group-level parameters may be successfully returned. The 95% bootstrap confidence intervals for each group term successfully uncover the simulation parameters, indicated by Table 3. Note that bolded and starred coefficients do not refer to statistically significant coefficients but whether the 95% confidence interval includes the corresponding simulation parameter. This indicates that the aforementioned routine allows users assess the generative structure for each role and return its underlying data generating process with a relatively high degree of precision. Not only does this technique allow the analyst to uncover unbiased estimates of each role's generative structure (conditional on the aforementioned assumptions), it provides the analyst coefficients that can be interpreted as TERGM parameters.

3 GOF Diagnostics

To assess the fit of the pooled TERGMs presented in the manuscript, goodness of fit (GOF) diagnostics are analyzed for each model (Hunter, Goodreau and Handcock 2008; Leifeld, Cranmer and Desmarais 2017). This section begins by discussing these diagnostics for the proof of concept model presented in Sections 1 and 2, then moves to a discussion of the GOF diagnostics for the pooled TERGMs in the Kapferer application.

3.1 Proof of Concept GOF

As a proof of concept, a pooled TERGM is fit on each set of networks assigned to a particular role. Given that the role generative process is known and properly specified, these TERGMs should fit particularly well. Figures 2 through 4 confirm this proposition, the observed statistics regularly intersect the median statistic values for the simulated networks.

3.2 Kapferer GOF

For the Kapferer strike network, a pooled TERGM is fit on the *In-Group* and *Out-Group* roles to assess their generative structure. The model fit for the *In-Group* appears to fit the observed data generating process reasonably well, as indicated in Figure 5. Broadly, the networks simulated reflect those observed with respect to the modularity of the network and the distribution of geodesic distances between nodes. While this is far from perfect, such close distributions for modularity would indicate a relatively well-fitting model. Different model specifications were attempted, and the only better fitting specification excluded GWESP, a theoretically-motivated term. The *Out-Group* TERGM fits relatively well, as indicated in Figure 6. Overall the model does quite well in approximating the overall structure of the network with respect to modularity and geodesic distance. While the observed statistics line does not intersect the median of the box-plots for the remaining statistics, the two trend together quite well.

[2]This table and all other tables presenting TERGM results were generated using Leifeld's `texreg` package (Leifeld 2013).

Figure 2 Goodness of Fit (GOF) Diagnostics for Role 1 TERGM. A well-fitting TERGM should have observed statistics (black line) that intersect the median statistic values values of the simulated networks (box-plots).

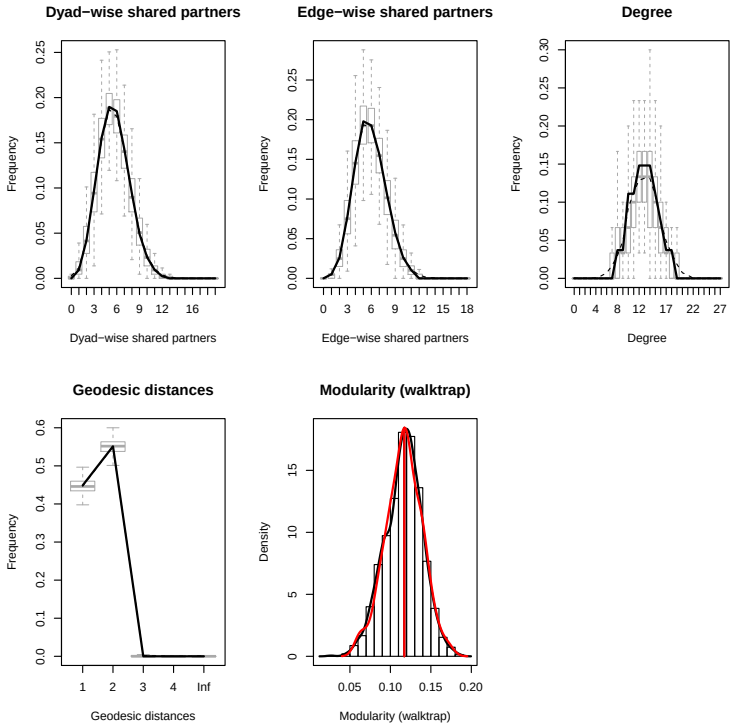


Figure 3 Goodness of Fit (GOF) Diagnostics for Role 2 TERGM. A well-fitting TERGM should have observed statistics (black line) that intersect the median statistic values values of the simulated networks (box-plots).

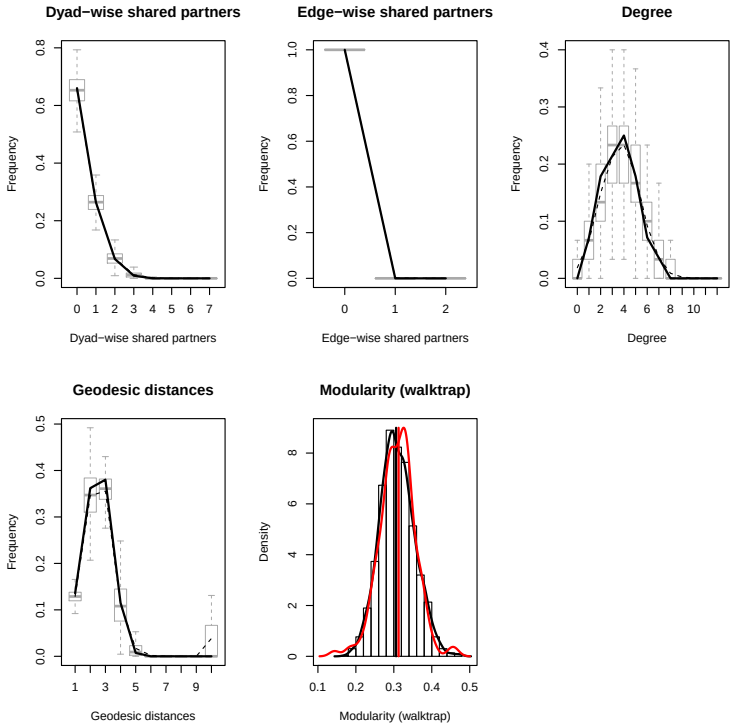


Figure 4 Goodness of Fit (GOF) Diagnostics for Role 3 TERGM. A well-fitting TERGM should have observed statistics (black line) that intersect the median statistic values values of the simulated networks (box-plots).

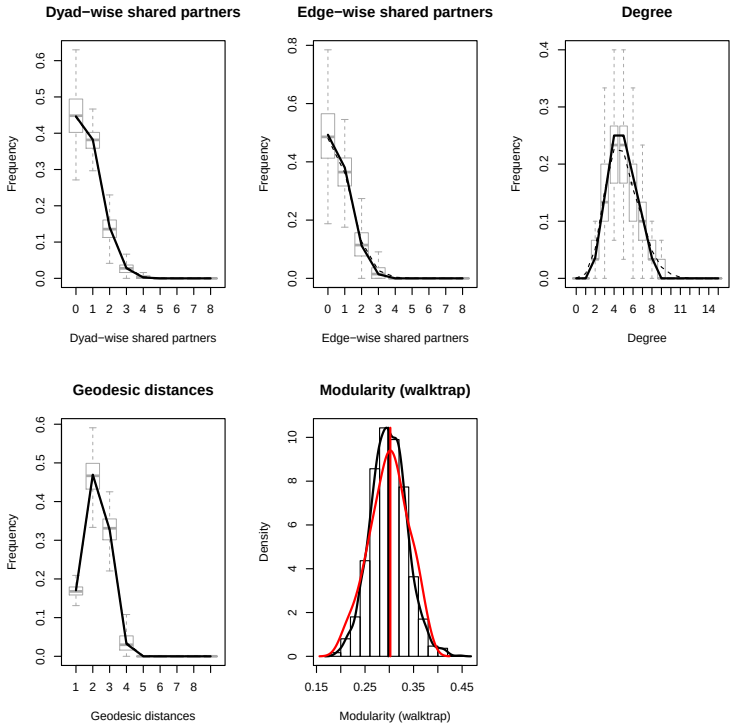


Figure 5 Goodness of Fit (GOF) Diagnostics for *In-Group* Role TERGM. A well-fitting TERGM should have observed statistics (black line) that intersect the median statistic values values of the simulated networks (box-plots).

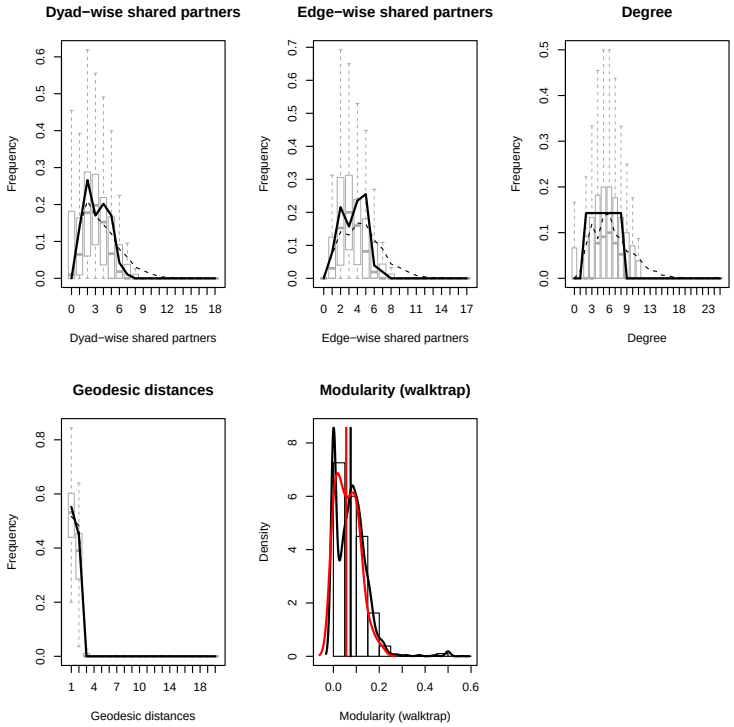
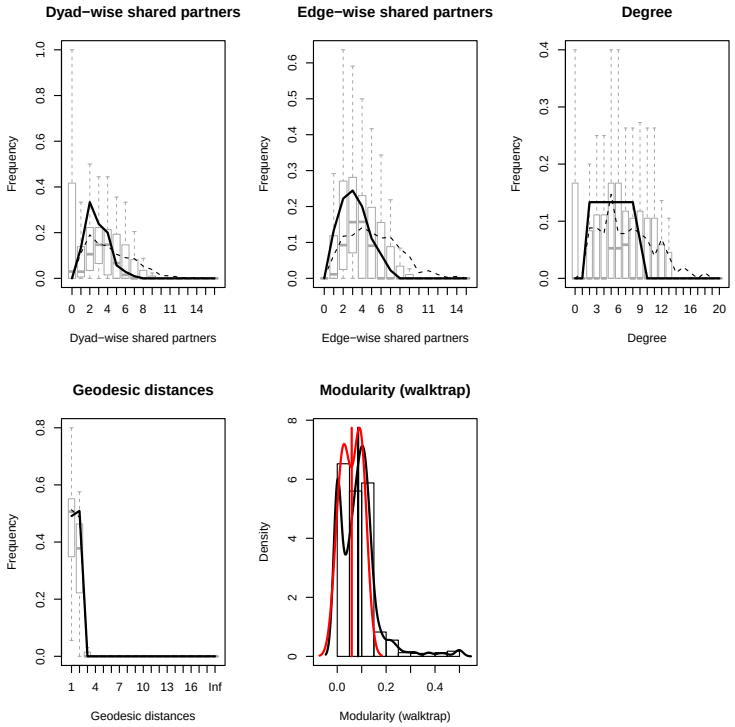


Figure 6 Goodness of Fit (GOF) Diagnostics for *Out-Group* Role TERGM. A well-fitting TERGM should have observed statistics (black line) that intersect the median statistic values values of the simulated networks (box-plots).



4 Ego-TERGM Pseudocode

To build intuition for the ego-TERGM's initialization and estimation procedure, I refer the reader to Algorithm 1. This algorithm presents pseudocode for the ego-TERGM. The algorithm is comprised of two core parts. The first is the two-step initialization process that fits a TERGM on each ego-network (n) and then clusters ego-networks according to similarity in model parameters. This produces initial cluster assignments $\mathbf{Z_{u=0}}$ and cluster centroids $\boldsymbol{\mu}$. Then, the EM algorithm proceeds in updating group assignments per step u until the change in log-likelihood ($loglikelihood_u - loglikelihood_{u-1}$) is less than the convergence parameter (α) or the number of potential steps ($steps$) has been exhausted.

Algorithm 1 Fitting procedure for ego-TERGM model.

```
N ← ego network list
G ← expected roles
H ← length(terms)
for n ∈ N do
    Calculate change statistics for n and its offset term ω
    Estimate TERGM via bootstrapped MPLE on ego-network n
    Save coefficients, change statistics, and offset terms
end for
Find G initial clusters of the length(N) × H matrix TERGM coefficients using k-means
Extract μ cluster centroids from k-means
Calculate length(N) × G matrix, Z_{u=0}, of cluster assignment probabilities
loglikelihood_{u=0} ← NaN
steps ← Maximum number of EM steps
θ_{u=0} ← μ
α ← convergence value
τ_{u=0} ← initial mixing proportions
for u ∈ steps do
    while loglikelihood_u − loglikelihood_{u−1} < α do
        for i ∈ length(N) do
            for g ∈ G do
                Update Z_u based upon θ_{u−1}, ω, and τ_{u−1} as per Equation 8
            end for
        end for
        Standardize Z_u for row sum of 1
        Update τ_u as the global mixing proportions
        Update loglikelihood_u using τ_u, Z_u, ω, and θ_{u−1} as per Equation 9
        Update θ_u using θ_{u−1}, Z_u, ω, and τ_u as per Equation 9
    end while
end for
```

References

Box-Steffensmeier, Janet M, Benjamin W Campbell, Dino P Christenson and Zachary Navabi. 2018. "Role Analysis Using the Ego-ERGM: A Look at Environmental Interest Group Coalitions." *Social Networks, Forthcoming* 52:213–227.

Brandes, U and J Lerner. 2007. Role equivalent Actors in Networks. In *Obiedkov S, Roth C. ICFCA Satellite Workshop on Social Network Analysis and Conceptual Structures: Exploring Opportunities*.

Cranmer, Skyler J and Bruce A Desmarais. 2011. "Inferential network analysis with exponential random graph models." *Political Analysis* 19(1):66–86.

Desmarais, Bruce A and Skyler J Cranmer. 2012. "Statistical mechanics of networks: Estimation and uncertainty." *Physica A: Statistical Mechanics and its Applications* 391(4):1865–1876.

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.

Leifeld, Philip. 2013. "texreg: Conversion of Statistical Model Output in R to LATEX and HTML Tables." *Journal of Statistical Software* 55(8):1–24.

Leifeld, Philip, Skyler J Cranmer and Bruce A Desmarais. 2017. "Temporal exponential random graph models with btergm: estimation and bootstrap confidence intervals." *Journal of Statistical Software* .

Salter-Townshend, Michael and Brendan Thomas Murphy. 2015. "Role analysis in networks using mixtures of exponential random graph models." *Journal of Computational and Graphical Statistics* 24(2):520–538.