

New network models facilitate analysis of biological networks

Alex Stivala

Università della Svizzera italiana, Via Giuseppe Buffi 13, 6900 Lugano, Switzerland
Email: alexander.stivala@usi.ch

December 12, 2023

Abstract

Exponential-family random graph models (ERGMs) are a family of network models originating in social network analysis, which have also been applied to biological networks. Advances in estimation algorithms have increased the practical scope of these models to larger networks, however it is still not always possible to estimate a model without encountering problems of model near-degeneracy, particularly if it is desired to use only simple model parameters, rather than more complex parameters designed to overcome the problem of near-degeneracy. Two new network models related to the ERGM, the Tapered ERGM, and the latent order logistic (LOLOG) model, have recently been proposed to overcome this problem. In this work I illustrate the application of the Tapered ERGM and the LOLOG to a set of biological networks, including protein-protein interaction (PPI) networks, gene regulatory networks, and neural networks. I find that the Tapered ERGM and the LOLOG are able to estimate models for networks for which it was not possible to estimate a conventional ERGM, and are able to do so using only simple model parameters. In the case of two neural networks where data on the spatial position of neurons is available, this allows the estimation of models including terms for spatial distance and triangle structures, allowing triangle motif statistical significance to be estimated while accounting for the effect of spatial proximity on connection probability. For some larger networks, however, Tapered ERGM and LOLOG estimation was not possible in practical time, while conventional ERGM models were able to be estimated only by using the Equilibrium Expectation (EE) algorithm.

Keywords— ERGM, exponential-family random graph model, LOLOG, latent order logistic model, biological networks, motif

1 Introduction

Networks are of great interest in biology in a variety of contexts, such as protein-protein interaction (PPI) networks (De Las Rivas and Fontanillo, 2010), gene regulatory networks (Emmert-Streib et al., 2014), and neural networks (Allard and Serrano, 2020). It is therefore important to have appropriate methods for analyzing such networks (de Silva and Stumpf, 2005), and as discussed in Stivala and Lomi (2021), the exponential-family random graph model (ERGM) is one such method, developed in the context of social networks (Lusher et al., 2013), and first applied to biological networks by Saul and Filkov (2007). Network models can be generative, seeking mechanisms and explanations, or null models, used for hypothesis testing (Betzler and Bassett, 2017). A generative model can, of course, also be used as a null model, by generating simulated networks and comparing the observed statistics of interest to their distributions in the simulated networks, as done with ERGM in, for example, Felmlee et al. (2021); Stivala and Lomi (2021).

Null models are often used for testing the statistical significance of “motifs”, small (usually connected and induced) subgraphs, which are thought to be over-represented (Milo et al., 2002; Shen-Orr et al., 2002). The conventional null model for assessing this significance is generated by randomizing the observed network, preserving the degree of each node (Milo et al., 2002), although more sophisticated randomization techniques, preserving additional properties, can also be used (Mahadevan et al., 2006; Orsini et al., 2015). The conventional procedure has intrinsic limitations, as described in Fodor et al. (2020), which can be overcome, at least in part, by using a more sophisticated model, such as the ERGM, as a generative model, a null model, or both (Stivala and Lomi, 2021). These problems can also be solved using an information theory approach, such as that recently described in Bénichou et al. (2023).

ERGMs are widely used for modeling social networks (Lusher et al., 2013; Koskinen, 2020, 2023), as well as other application fields (Ghafouri and Khasteh, 2020), and development of ERGM modelling techniques is an active

research area (Cimini et al., 2019; Lusher et al., 2020; Schweinberger et al., 2020; Schmid et al., 2022; Stivala and Lomi, 2022; Krivitsky et al., 2022, 2023; Caimo and Gollini, 2023; Giacomarra et al., 2023; Schmid and Hunter, 2024). A brief survey of ERGM applications to biological networks is given in Stivala and Lomi (2021), but some notable more recent work on the application of ERGMs to neural networks includes the application of temporal ERGMs (Leifeld et al., 2018) to model brain network reorganization after stroke (Obando et al., 2022), a detailed review of ERGMs in brain connectivity networks (Dichio and De Vico Fallani, 2023b), and the use of ERGMs as part of a novel method to explore development of the *C. elegans* connectome (Dichio and De Vico Fallani, 2023a). The *C. elegans* neural network has also been the subject of a related modeling technique, stochastic blockmodelling (Pavlovic et al., 2014; Gross et al., 2023).

As discussed in Stivala and Lomi (2021), for some of the biological networks considered it was not possible to estimate an ERGM, and in particular neural networks proved particularly problematic. In this work, I use two new, related, models, the Tapered ERGM (Fellows and Handcock, 2017; Blackburn and Handcock, 2023) and the latent order logistic model (LOLOG) (Fellows, 2018; Clark and Handcock, 2022) to re-analyze the biological networks considered in Stivala and Lomi (2021) and compare the results to the ERGMs presented there, as well as to analyze some neural networks for which ERGM models could not be estimated.

2 Network models

2.1 Exponential-family random graph model (ERGM) and Tapered ERGM

An ERGM is a probability distribution with the form

$$p_{\text{ergm}}(x | \theta) = \frac{1}{\kappa(\theta)} \exp \left(\sum_A \theta_A g_A(x) \right) \quad (1)$$

where

- $X = [X_{ij}]$ is square binary matrix of random tie variables,
- x is a realization of X ,
- A is a “configuration”, a (small) set of nodes and a subset of ties between them,
- $g_A(x)$ is the network statistic for configuration A ,
- θ_A is a model parameter corresponding to configuration A ,
- $\kappa(\theta) = \sum_{x \in G_N} \exp(\sum_A \theta_A g_A(x))$, where G_N is the set of all square binary matrices of order N (graphs with N nodes), is the normalizing constant to ensure a proper distribution.

Configurations can include nodal attributes, including dyadic attributes such as, for example, the distance between each pair of nodes. These are assumed to be fixed (exogenous to the model).

By estimating the parameter vector θ which maximizes the probability of an observed graph x under the model, that is, the maximum likelihood estimate (MLE), together with the estimated standard errors of the parameter estimates, inferences can be made about the over-representation (a positive and statistically significant parameter estimate) or under-representation (a negative and statistically significant parameter estimate) of the corresponding configurations. These inferences about over-representation (or under-representation) are, for each configuration, conditional on all the other configurations included in the model, which need not be independent (and which, indeed are usually not independent). For example, a model might include parameters for density, degree distribution, triangles, two-paths (note that a triangle is formed by adding a third edge to a two-path), and spatial distance between nodes. In this model, a statistically significant positive estimate for triangles would indicate that the triangle configuration occurs more frequently than expected by chance, even accounting for density, degree distribution, number of two-paths, and the effect of spatial distance on edge probability.

ERGM parameter estimation is, due to the intractable normalizing constant $\kappa(\theta)$ in (1), a computationally intractable problem, and hence in practice it is necessary (except for extremely small networks, (Vega Yon et al., 2021)) to use Markov chain Monte Carlo (MCMC) methods (Hunter et al., 2012) such as Markov Chain Monte Carlo MLE

(MCMCMLE) (Geyer and Thompson, 1992) or stochastic approximation (Snijders, 2002). More recently, the ‘‘Equilibrium Expectation’’ (EE) algorithm (Byshkin et al., 2016, 2018; Borisenko et al., 2020) has allowed ERGM models for networks, including biological networks (Byshkin et al., 2018; Stivala and Lomi, 2021), far larger than previously possible, to be estimated in practical time (Stivala et al., 2020). Note that Stivala et al. (2020) states that, ‘‘An important next step is the strengthening of the theoretical basis for the EE algorithm . . . there are no theoretical guarantees behind the EE algorithm’’ (Stivala et al., 2020, p. 17). However that referred to the original EE algorithm (Byshkin et al., 2018), rather than the simplified EE algorithm (Borisenko et al., 2020) now (also) implemented in `EstimNetDirected`.^{1,2} This algorithm has been shown to converge to the MLE, if it exists, when the learning rate is small enough, by Giacomarra et al. (2023), using the uncertain energies framework of Ceperley and Dewing (1999), a proof first outlined briefly (but not for the specific case of connected networks considered in Giacomarra et al. (2023)) by Borisenko et al. (2020) using the results of Ceperley and Dewing (1999); Frenkel et al. (2017).

It is a well-known problem with ERGMs that simple model specifications (such as a model containing only terms for edges and triangles) can result in ‘‘near-degeneracy’’, where the MLE does not exist, or the probability mass is concentrated in a small subset of graphs, often (nearly) empty or (nearly) complete graphs (Handcock, 2003; Snijders et al., 2006; Hunter, 2007; Schweinberger, 2011; Chatterjee and Diaconis, 2013; Schweinberger et al., 2020; Blackburn and Handcock, 2023). The usual solution to this problem is to use ‘‘alternating’’ or ‘‘geometrically weighted’’ configurations (Snijders et al., 2006; Robins et al., 2007; Hunter, 2007; Lusher et al., 2013; Stivala, 2023), however this can result in the model no longer directly testing the hypotheses of interest (Stivala and Lomi, 2021; Blackburn and Handcock, 2023), or researchers omitting model terms due only to their tendency to near-degeneracy, or omitting models entirely, due to an inability to obtain converged estimates (Martin, 2017, 2020; Clark and Handcock, 2022).

Tapered ERGMs (Fellows and Handcock, 2017; Blackburn and Handcock, 2023) are a variant of the ERGM that solves the problem of near-degeneracy by imposing upper bounds on the variance of the sufficient statistics. The Tapered ERGM works by assigning lower probability to graphs with statistics far from their central location, using a new tapering parameter vector τ . The parameter vector τ can itself be estimated by optimizing a double penalized likelihood, penalizing kurtosis values too far from the target values and values of τ that are too large (Blackburn and Handcock, 2023). The Tapered ERGM is implemented in the `ergm.tapered` R package (Handcock et al., 2022b,a; Krivitsky et al., 2022, 2023).

2.2 Latent order logistic (LOLOG) model

The latent order logistic (LOLOG) model (Fellows, 2018; Clark and Handcock, 2022) is related to the ERGM, but is based on the principle of network growth, and in particular, a (latent) node ordering process. Each edge variable is sequentially considered for creation, and edges are not deleted.

In the following, X is a random graph with N nodes ($[X_{ij}]$ is square binary matrix of random tie variables), x is a realization of X , and X_t , $t = 1, \dots, |x|$ are latent random variables representing the sequential formation of X . X_t has exactly t edges and is formed from X_{t-1} by the addition of a single edge (Fellows, 2018; Clark and Handcock, 2022). Here $|x|$ is the number of dyads in the graph, and hence for directed graphs $|x| = N(N - 1)$ and for undirected graphs $|x| = N(N - 1)/2$.

A LOLOG model is specified by two components (Fellows, 2018; Clark and Handcock, 2022). First, the probability of observing a graph given a specified edge formation order s , which is a product of logistic likelihoods:

$$p(x | s, \theta) = \prod_{t=1}^{|x|} \frac{1}{Z_t(s)} \exp(\theta \cdot C_{s,t}) \quad (2)$$

where

- $s = \{s_1, s_2, \dots, s_{|x|}\}$ is the set of all possible edge formation orders with $|x|$ dyads,
- $C_{s,t} = g(x_t, s_{\leq t}) - g(x_{t-1}, s_{\leq t-1})$ where $s_{\leq t}$ denotes the first t elements of s and $g(\cdot)$ are the sufficient statistics; $C_{s,t}$ are the change statistics,

¹<https://github.com/stivalaa/EstimNetDirected>

²The simplified EE algorithm of Borisenko et al. (2020) is used if `useBorisenkoUpdate = true` is set in the `EstimNetDirected` configuration file, otherwise the original EE algorithm described in Byshkin et al. (2018); Stivala et al. (2020) is used. I recommend always setting `useBorisenkoUpdate = true`, as was done, for example, for estimating the ERGM models described in Stivala and Lomi (2021, 2022); Stivala et al. (2023b).

- $Z_t(s) = \exp(g(x_t^+, s_{\leq t}) - g(x_{t-1}, s_{\leq t-1})) + 1$, where x_t^+ is the graph x_{t-1} with edge s_t added, are the normalizing constants.

The second component is the marginal likelihood of an observed graph, a sum over all possible edge permutations, where $p(s)$ is the probability that the ordering s occurs:

$$p_{\text{lolog}}(x | \theta) = \sum_s p(x | s, \theta) p(s) \quad (3)$$

LOLOG model estimation by Monte Carlo method of moments is implemented in the lolog R package (Fellows, 2019). If a partial edge ordering is observed (based on the order of nodes being added to the network, for example), that the edge ordering can be constrained according to this data, otherwise the space of all possible edge permutations is randomly sampled.

It is suggested in Clark and Handcock (2022, p. 571) that the reason that the LOLOG does not suffer from the near-degeneracy problem that ERGM models do, is that, unlike ERGM, the LOLOG simulation procedure considers each dyad exactly once, limiting the scope for “explosive” edge formation (or dissolution, which does not occur at all in the LOLOG simulation procedure) that can occur in ERGM.

Clark and Handcock (2022), by re-analyzing a substantial set of networks first analyzed with the ERGM, find that the LOLOG is generally in qualitative agreement with the ERGM, and is often able to fit with simpler terms that would result in near-degeneracy in ERGM. Further, they find that the LOLOG tends to be easier (and faster) to fit.

As a recently developed model, there are far fewer published uses of the LOLOG than there are of the well-established ERGM. However, as well as the re-analyses of networks originally modelled with ERGMs in Clark and Handcock (2022), some applications to date include analysis of YouTube video recommendations (Abul-Fottouh et al., 2020; Gruzd et al., 2023) and boilerplate language in international trade agreements (Peacock et al., 2019) and environmental impact statements (Scott et al., 2021).

3 Methods

3.1 Network data

Seven biological networks are examined in this work. Four of them are exactly those used in Stivala and Lomi (2021): two undirected PPI networks, and two directed gene regulatory networks. The remaining three are neural networks.

The *Saccharomyces cerevisiae* (yeast) PPI network (von Mering et al., 2002), HIPPIE human PPI network (Schaefer et al., 2012, 2013; Suratanee et al., 2014; Alanis-Lobato et al., 2016), *Alon E. coli* regulatory network (Salgado et al., 2001; Shen-Orr et al., 2002), and yeast regulatory network (Milo et al., 2002; Costanzo et al., 2001) are exactly those used in Stivala and Lomi (2021), and their data processing is described in detail there.

The first neural network is the whole-animal chemical connectome of the adult male *C. elegans* worm (Cook et al., 2019). This was mentioned in Stivala and Lomi (2021) as a network for which a converged ERGM could not be found. The network was obtained from the *C. elegans* connectome tables in MATLAB-ready format (corrected July 2020) by Kamal Premaratne (University of Miami) downloaded from <https://wormwiring.org/matlab%20scripts/Premaratne%20MATLAB-ready%20files%20.zip> (accessed 6 September, 2021). The data was converted using the R.matlab (Bengtsson, 2018) and igraph (Csárdi and Nepusz, 2006) packages in R.

The second neural network is the hermaphrodite *C. elegans* neural network of 277 neurons. This data, including the spatial two-dimensional positions in the lateral plane (in mm) of the neurons (Choe et al., 2004; Kaiser and Hilgetag, 2006), and the birth times (in minutes) of the neurons (Sulston and Horvitz, 1977; Sulston et al., 1983; Varier and Kaiser, 2011) was downloaded from https://www.dynamic-connectome.org/?page_id=25 (accessed 31 May 2019).³ There was no birth time data for the hermaphrodite-specific ventral chord motor neuron VC6, which was excluded from the analysis in Varier and Kaiser (2011), so this was imputed as the mean birth time of the other VCn neurons.

The third neural network is the *Drosophila* optic medulla connectome (Takemura et al., 2013), obtained via the Open Connectome database (Vogelstein et al., 2018) at <http://openconnectome.me/graph-services/download/> (accessed 6 November 2016).⁴ Following Agarwala and Kenter (2023), three-dimensional geometric locations were assigned to the neurons using the centroids of the synaptic coordinates associated with them, and the

³This data is now available from <https://sites.google.com/view/dynamicconnectomelab/resources>.

⁴This data is now available from <https://neurodata.io/data/takemura13/>.

subgraph induced by the “named” nodes, consisting of those neurons that have alphanumeric labels, was extracted. As described in Takemura et al. (2013); Agarwala and Kenter (2023), the neighbourhoods of the 379 neurons in the central neural column of the *Drosophila* medulla were traced, but not necessarily the connections between the other neurons, and the 358 “named” neurons “correspond roughly to the central 379 nodes” (Agarwala and Kenter, 2023). Unlike Agarwala and Kenter (2023), however, I do not treat the network as undirected, and do not remove the highest degree node (this is necessary in Agarwala and Kenter (2023) as it violates a condition of their model). I also do not consider the full network, but only the network induced by the named nodes, since only these have their connections fully traced.

For all networks, multiple edges and loops (an edge or arc from a node to itself) are removed. In the case of the *Drosophila* medulla, this is another difference from Agarwala and Kenter (2023), where loops are retained. In the *E. coli* network, following Hummel et al. (2012); Stivala and Lomi (2021), where the self-loops indicate self-regulation, this is represented instead by a binary node attribute “self”, which is true when a self-loop was present, and false otherwise.

Summary statistics of the networks, computed using the igraph (Csárdi and Nepusz, 2006) R package, are shown in Table 1.

Table 1: Network summary statistics.

Network	Directed	N	Components	Size of largest component	Mean degree	Density	Clustering coefficient
Yeast PPI	N	2617	92	2375	9.06	0.00346	0.46862
Human PPI (HIPPIE)	N	11517	93	11322	8.19	0.00071	0.03773
Alon <i>E. coli</i> regulatory	Y	423	34	328	2.45	0.00291	0.02382
Alon yeast regulatory	Y	688	11	662	3.14	0.00228	0.01625
Cook <i>C. elegans</i> connectome	Y	575	17	559	18.25	0.01589	0.25776
Kaiser <i>C. elegans</i> neural	Y	277	1	277	15.20	0.02753	0.19809
<i>Drosophila</i> medulla (named)	Y	358	8	351	20.07	0.02811	0.17798

“Clustering coefficient” is the global clustering coefficient (transitivity).

The in-degree and out-degree distributions of the neural networks are shown in Figure 1 as plots of their empirical cumulative distribution function (CDF). In this figure, α is the exponent in the discrete power law distribution $\Pr(X = x) = Cx^{-\alpha}$ (where C is a normalization constant), and μ and σ are the parameters (respectively, mean and standard deviation of $\log(x)$) of the discrete log-normal distribution. Discrete power law and log-normal distributions were fitted using the methods of Clauset et al. (2009) implemented in the poweRlaw package (Gillespie, 2015). Degree distributions of the other networks are shown in Stivala and Lomi (2021, Fig.3).

For the Cook *C. elegans* connectome, for both log-normal and power law, and for both in-degree and out-degree distributions, the null hypothesis that the tail of the empirical distribution ($x_{\min} = 11$ for in-degree for both power law and log-normal, and for out-degree, $x_{\min} = 26$ for power law and $x_{\min} = 3$ for log-normal) is consistent with the log-normal or power law distribution cannot be rejected at the conventional $p < 0.05$ level.

For the Kaiser *C. elegans* neural network, the null hypotheses that the tails of the in-degree and out-degree distributions are consistent with power law ($x_{\min} = 15$ for in-degree, $x_{\min} = 10$ for out-degree) or log-normal distributions ($x_{\min} = 5$ for both in-degree and out-degree distributions) cannot be rejected at the conventional $p < 0.05$ level.

For the *Drosophila* optic medulla network, for the in-degree distribution, the null hypothesis that the tail of the empirical CDF ($x_{\min} = 13$) is consistent with a power law distribution is rejected ($p < 0.05$), while for the log-normal distribution ($x_{\min} = 4$) the null hypothesis is not rejected at the conventional $p < 0.05$ level. For the out-degree distribution, neither of the null hypotheses that the tail of the empirical CDF is consistent with a power law distribution ($x_{\min} = 26$) or with a log-normal distribution ($x_{\min} = 6$) can be rejected at the conventional $p < 0.05$ level.

3.2 Model estimation

All estimations, unless otherwise noted, were run using R (R Core Team, 2022, version 4.0.2) packages on an Intel Xeon E5-2650 v3 2.30 GHz CPU on a Linux compute cluster node. Facilitating direct comparisons of estimation time, this is the same cluster that was used for ERGM estimations using the Estimnet (Byshkin et al., 2016, 2018), EstimNetDirected (Stivala et al., 2020), and statnet (Handcock et al., 2008, 2016) ergm software in Stivala and Lomi

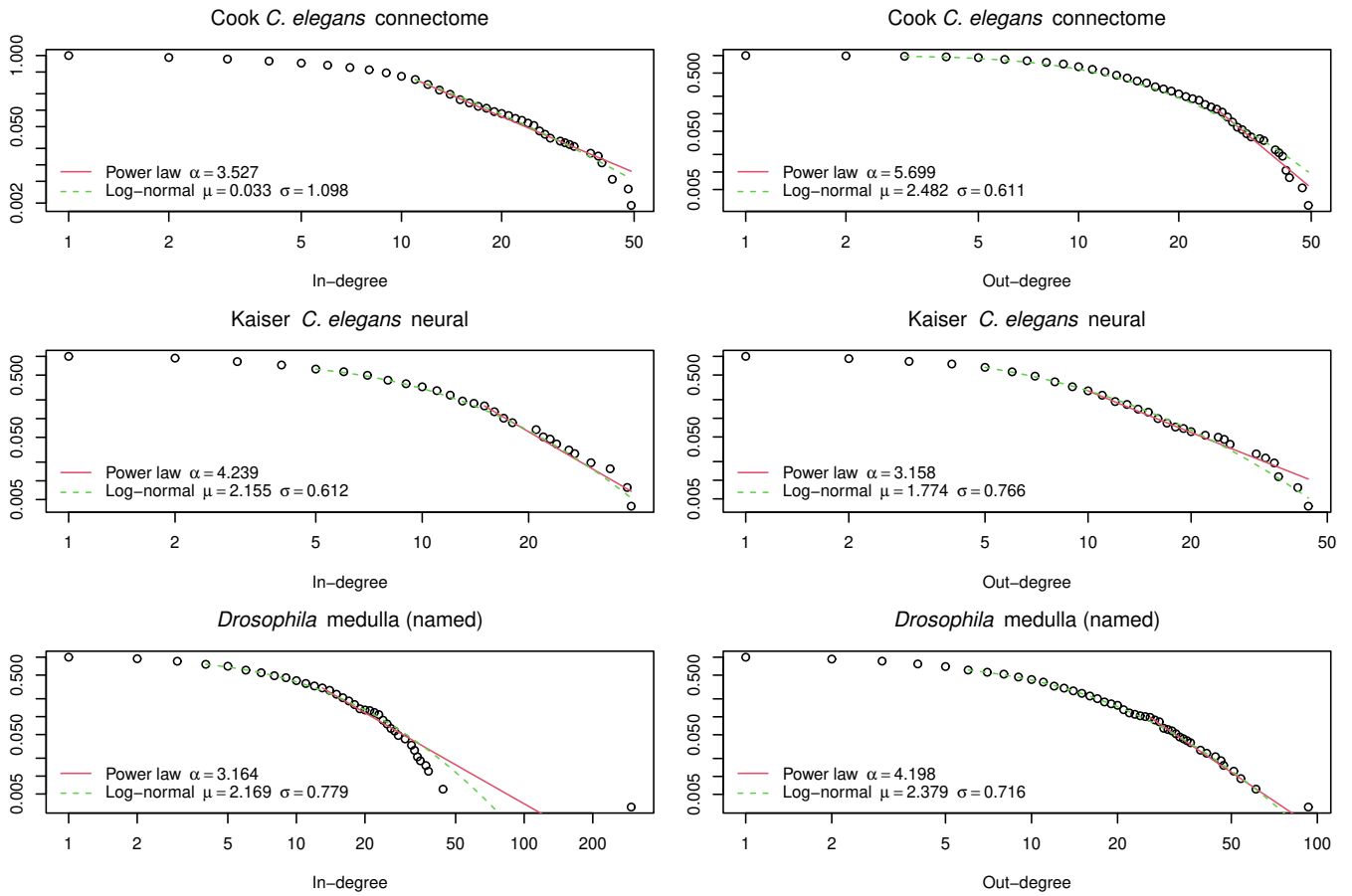


Figure 1: Discrete power law and log-normal distributions fitted to the empirical CDFs for in-degree (left) and out-degree (right) distributions of the neural networks.

(2021). Note that these programs implement stochastic (Monte Carlo) algorithms and terminate when certain convergence criteria are met, and so can demonstrate a large variance in runtime. Hence the reported times, for a single run for each of the estimations shown here, are merely indicative of the approximate computational requirements, and are not necessarily representative.

For LOLOG estimations, the lolog R package (Fellows, 2019) was used, and for Tapered ERGM estimations, the `ergm.tapered` R package (Handcock et al., 2022b) was used. The latter package builds on the `statnet` (Handcock et al., 2008, 2016) `ergm` package (Handcock et al., 2022a; Krivitsky et al., 2022, 2023). No parallel computing was used (each estimation was run on a single core), and a 48 hour maximum time limit was imposed by the cluster job management system.

For estimating Tapered ERGMs with the `ergm.tapered` package, tapering for dependent parameters was estimated using kurtosis-penalized likelihood, as described in Blackburn and Handcock (2023). For geometrically weighted LOLOG or ERGM terms, the decay parameter, denoted α in models presented here, was fixed at values based on those used in Stivala and Lomi (2021), and adjusted if necessary by trial and error to improve model convergence and fit, as estimating the decay parameter is not implemented in LOLOG, and curved ERGM model estimation (Hunter and Handcock, 2006; Hunter, 2007) resulted in estimation not converging within the time limit. LOLOG models were estimated with default estimation parameters, and model convergence was confirmed by inspecting the model diagnostic plots (shown in Appendix A). Tapered ERGM models were estimated with `ergm.tapered` with default estimation parameters, using the `ergm` version 4 package (Handcock et al., 2022a) adaptive MCMC via effective sample size feature to automatically adjust the required number of iterations (Krivitsky et al., 2022). Model convergence was confirmed by inspecting the MCMC diagnostic plots, and only converged non-degenerate models were used.

4 Results and discussion

4.1 Protein-protein interaction (PPI) networks

Yeast PPI and human PPI (HIPPIE) LOLOG estimations, and Tapered ERGM estimations of the HIPPIE network, did not converge within the 48 hour maximum time limit. Hence the only PPI network model I was able to estimate was a Tapered ERGM model for the yeast PPI network (Table 2). This estimation took approximately 9 hours.

Table 2: Tapered ERGM (Fellows and Handcock, 2017; Blackburn and Handcock, 2023) parameter estimates (with estimated standard errors in parentheses) for the yeast PPI network. Estimated using the `ergm.tapered` package (Handcock et al., 2022b).

Effect	Model 1
Edges	-6.215 (0.013) ^{***}
GW degree ($\alpha = 0.1$)	-2.478 (0.125) ^{***}
Two-paths	0.015 (0.002) ^{***}
Triangles	0.100 (0.005) ^{***}
AIC	154767.00
BIC	154832.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; \cdot $p < 0.1$.

The model in Table 2 shows that triangles are over-represented in this network (the Triangles parameter is positive and statistically significant). The “geometrically weighted degree” parameter is negative and significant, indicating centralization of edges (Hunter, 2007; Levy, 2016; Levy et al., 2016). These results are consistent with the ERGM model of Stivala and Lomi (2021, Table 4), in which there are a positive and significant estimated parameters for both the “alternating k -triangle“ and “alternating k -star” effects. Note that, confusingly (Levy et al., 2016; Martin, 2020; Stivala, 2020), the interpretation of the geometrically weighted degree (GW degree) parameter defined in Hunter (2007) and used in `statnet` (and hence the tapered ERGM in Table 2) has a different interpretation regarding the sign to that of the “alternating k -star” effect (Snijders et al., 2006; Robins et al., 2007) used in Stivala and Lomi (2021). A positive value of the alternating k -star effect indicates centralization based on high-degree nodes, a situation which is indicated by a negative value of the geometrically-weighted degree parameter.

Goodness-of-fit plots for this model are shown in Figure A1, showing that the fit to the triad census (and hence

number of triangles) is excellent. However the fit to the other network statistics, and in particular the geodesic distance distribution, is poor.

Using the tapered ERGM, therefore, allowed estimation of a model more directly testing motif significance (using a parameter for triangles directly, rather than requiring alternating k -triangles or “geometrically weighted edgewise shared partners”). However I was not able to estimate a Tapered ERGM model for the larger human PPI network, nor a LOLOG model for either of the PPI networks. ERGM models for these networks could also not be estimated with “standard” ERGM estimation software, specifically PNet (Wang et al., 2009), MPNet (Wang et al., 2014, 2022), or the statnet ergm package (Handcock et al., 2022a; Krivitsky et al., 2023). In contrast, ERGM models for both networks could be estimated using the “Equilibrium Expectation” (EE) algorithm (Byshkin et al., 2018; Borisenko et al., 2020) as described in Stivala and Lomi (2021).

It is also worth noting that, while the “alternating k -two-paths” parameter, when included, was found to be negative and significant in Stivala and Lomi (2021, Table 4), in the different Tapered ERGM model (in which simple Triangles and Two-paths are used, rather than “alternating” versions) shown in Table 2, the Two-paths parameter is found to be positive and significant. No interpretation of these parameters is made, but they are included to “control for” the triangle (or alternating k -triangle) parameters, since, as discussed in Stivala and Lomi (2021), these ERGM configurations are not *induced* subgraphs. That is, a triangle contains two-paths within it as a subgraph (but not an induced subgraph), and so it is usual ERGM modeling practice to include a term for two-paths when a term for triangles is included (Koskinen and Daraganova, 2013).

One further point to note, is that in Stivala and Lomi (2021), a nodal attribute for the functional category (class) of the protein was used, and a parameter to test for matching class included (Stivala and Lomi, 2021, Table 4). However a significant number of proteins are “uncharacterized” (have an NA value for class) (von Mering et al., 2002, S. I.), and NA values for nodal attributes are not currently allowed for Tapered ERGM models in statnet; some form of imputation or other technique for handling missing data (Koskinen et al., 2013) would have to be used.

4.2 Gene regulatory networks

Table 3: Latent order logistic (LOLOG) model (Fellows, 2018) parameter estimates for the Alon *E. coli* regulatory network. Estimated using the lollog package (Fellows, 2019).

Effect	Estimate	Std. error	p-value
Edges	-5.9596	0.5955	< 0.001
GW in-degree ($\alpha = 0.2$)	-2.7560	0.1564	< 0.001
GW out-degree ($\alpha = 0.2$)	1.5308	0.2016	< 0.001
Two-paths	-0.8845	0.1564	< 0.001
Triangles	2.9323	0.2463	< 0.001
Nodecov self	0.7223	0.2084	< 0.001
Nodematch self	-0.8923	0.1420	< 0.001

A LOLOG model for the *E. coli* gene regulatory network is shown in Table 3. This estimation took approximately 20 minutes. Tapered ERGM models for the network are shown in Table 4. Estimation of Model 1 and Model 2 took approximately 1 minute and 33 minutes, respectively. Consistent with the ERGM models for this network in Stivala and Lomi (2021), both the LOLOG and Tapered ERGM models indicate centralization based on high-in-degree nodes, but not (indeed, a tendency against such centralization) for out-degree, an over-representation of transitive triangles, and a tendency against self-regulating operons regulating other self-regulating operons (negative and significant estimate for the “Nodematch self” parameter).

This is the only network examined here in which an ERGM was able to be estimated with “standard” statnet ergm package (Stivala and Lomi, 2021, Table S1), rather than requiring the use of the EE algorithm in EstimNetDirected. The LOLOG and Tapered ERGM models shown here were able to be estimated with the Triangles or Transitive triples, respectively, parameter, rather than requiring the alternating k -triangles or geometrically weighted edgewise shared partners (GWESP) parameter used in Stivala and Lomi (2021), however. The LOLOG and statnet (and hence the Tapered ERGM), packages, however, do not allowing the modelling of loops (self-edges), and hence the nodal covariate “self” is used here, as in Hummel et al. (2012) to model self-regulation in a simplistic way. In contrast, EstimNetDirected can model networks containing loops, and it was found that loops are over-represented in this

Table 4: Tapered ERGM (Fellows and Handcock, 2017; Blackburn and Handcock, 2023) parameter estimates (with estimated standard errors in parentheses) for the Alon *E. coli* regulatory network. Estimated using the `ergm.tapered` package (Handcock et al., 2022b).

Effect	Model 1	Model 2
Edges	-3.062 (0.077) ^{***}	-2.716 (0.156) ^{***}
GW in-degree ($\alpha = 2$)	-3.979 (0.134) ^{***}	-3.831 (0.134) ^{***}
GW out-degree ($\alpha = 0$)	1.570 (0.190) ^{***}	1.479 (0.188) ^{***}
Two-paths	-0.495 (0.102) ^{***}	-0.501 (0.101) ^{***}
Transitive triples	2.171 (0.234) ^{***}	2.156 (0.177) ^{***}
Nodecov self		-0.230 (0.122)
Nodematch self		-0.524 (0.139) ^{***}
AIC	5656.00	5635.00
BIC	5717.00	5715.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.1$.

network (Stivala and Lomi, 2021).

Goodness-of-fit plots for the LOLOG (Table 3) and Tapered ERGM (Table 4) are shown Figure A2 and Figure A3, respectively, showing all models fit the data well.

Table 5: Latent order logistic (LOLOG) model (Fellows, 2018) parameter estimates for the Alon yeast regulatory network. Estimated using the `lolog` package (Fellows, 2019).

Effect	Estimate	Std. error	p-value
Edges	-4.1067	0.0777	< 0.001
GW in-degree ($\alpha = 0.2$)	1.0386	0.1048	< 0.001
GW out-degree ($\alpha = 0.2$)	-3.9870	0.1178	< 0.001
Two-paths	-0.7062	0.0656	< 0.001
Triangles	2.6633	0.1659	< 0.001

A LOLOG model for the yeast regulatory network is shown in Table 5. This estimation took approximately 36 minutes. The Tapered ERGM estimation of this network shown in Table 6 took approximately 5 minutes. Consistent with the results in Stivala and Lomi (2021, Table 7), these models show centralization on out-degree, but not in-degree, and an over-representation of (transitive) triangles.

Goodness-of-fit plots for the LOLOG model (Table 5) and Tapered ERGM model (Table 6) are shown in Figure A4 and Figure A5, respectively. These figures show a good fit to the data, with the exception of the reciprocity (mutual) statistic for the tapered ERGM model (Table 6, Fig. A5). As discussed in Stivala and Lomi (2021), there is only a single reciprocated arc in this network, and hence ERGM models that do not include the Reciprocity parameter can show poor goodness-of-fit on statistics involving reciprocated arcs, while models that do include this parameter can exhibit poor convergence. I also found this to be the case with the Tapered ERGM, with the model shown in Table 6 showing good convergence diagnostics, while a model (not shown) without the Reciprocity (mutual) parameter does not. This issue does not affect the LOLOG model with no Reciprocity (mutual) parameter (Table 5), however, where the goodness-of-fit for the mutual statistic is good (Fig. A4).

4.3 Neural networks

A LOLOG model for the Cook *C. elegans* connectome is shown in Table 7. This estimation took approximately 3 hours. Tapered ERGM models for this network are shown in Table 8. Estimation of Model 1 and Model 2 took approximately 3 hours, and 7 hours, respectively.

These models both show centralization based on high out-degree nodes, and an over-representation of (transitive) triangles. The Tapered ERGM models also have a negative and statistically significant GW in-degree parameter, indicating centralization also for the in-degree distribution, but this parameter is not statistically significant in the LOLOG model. Conversely, the LOLOG model has a negative and statistically significant Two-paths parameter, but

Table 6: Tapered ERGM (Fellows and Handcock, 2017; Blackburn and Handcock, 2023) parameter estimates (with estimated standard errors in parentheses) for the Alon yeast regulatory network. Estimated using the `ergm.tapered` package (Handcock et al., 2022b).

Effect	Model 1
Edges	-3.858 (0.067) ^{***}
Reciprocity	-4.707 (2.747)
GW in-degree ($\alpha = 0.5$)	1.112 (0.170) ^{***}
GW out-degree ($\alpha = 1.5$)	-4.540 (0.108) ^{***}
Two-paths	-0.291 (0.065) ^{***}
Transitive triples	1.815 (0.145) ^{***}
AIC	11964.00
BIC	12041.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.1$.

Table 7: Latent order logistic (LOLOG) model (Fellows, 2018) parameter estimates for the Cook *C. elegans* connectome. Estimated using the `lolog` package (Fellows, 2019).

Effect	Estimate	Std. error	p-value
Edges	-2.8910	0.1179	< 0.001
GW in-degree ($\alpha = 0.1$)	0.3020	0.1874	0.1071
GW out-degree ($\alpha = 0.1$)	-3.9835	0.2112	< 0.001
Two-paths	-0.4339	0.0317	< 0.001
Triangles	3.6632	0.2081	< 0.001

neither Two-paths (Model 1) nor geometrically-weighted dyadwise shared partners (GWDSF, Model 2) are significant in the Tapered ERGM model (Table 8).

Goodness-of-fit plots for the LOLOG model (Table 7) and Tapered ERGM models (Table 8) are shown in Figure A6 and Figure A7, respectively. The LOLOG model has a poor fit on reciprocity (mutual) and the in-degree distribution, but a better fit on the out-degree distribution and edgewise shared partners. I also estimated a model (not shown) including the mutual (Reciprocity) parameter, however this model showed poor convergence on the diagnostic plots. The Tapered ERGM models show not very good fit on the degree distributions, and bad fit on the dyadwise and edgewise shared partner and geodesic distance distributions, but reasonable fit on the triad census.

As discussed in Stivala and Lomi (2021), a converged ERGM could not be found for this network using `statnet` (`ergm`) or `EstimNetDirected`, and hence this is an example where the LOLOG and Tapered ERGM can both estimate models that could not be estimated with “standard” ERGMs. This network, as well as the Kaiser *C. elegans* and

Table 8: Tapered ERGM (Fellows and Handcock, 2017; Blackburn and Handcock, 2023) parameter estimates (with estimated standard errors in parentheses) for the Cook *C. elegans* connectome. Estimated using the `ergm.tapered` package (Handcock et al., 2022b).

Effect	Model 1	Model 2
Edges	-3.860 (0.027) ^{***}	-3.747 (0.029) ^{***}
GW in-degree ($\alpha = 0.2$)	-4.458 (0.184) ^{***}	-4.499 (0.189) ^{***}
GW out-degree ($\alpha = 0.2$)	-9.269 (0.164) ^{***}	-9.381 (0.164) ^{***}
Two-paths	0.001 (0.003)	
GWDSF ($\alpha = 1$)		-0.006 (0.005)
Transitive triples	0.101 (0.005) ^{***}	0.102 (0.008) ^{***}
AIC	50449.00	50515.00
BIC	50514.00	50580.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.1$.

Drosophila medulla neural networks, for which I also could not find good ERGM models with statnet (ergm) or EstimNetDirected, although not larger than the other, non-neural, networks, have far higher mean degree and density (Table 1), which might be one reason contributing to my inability to find a converged model for such networks with other ERGM estimation algorithms.

Table 9: Latent order logistic (LOLOG) model (Fellows, 2018) parameter estimates for the Kaiser *C. elegans* neural network. Model 1 has node inclusion order specified by birth time, while Model 2 does not. Estimated using the lollog package (Fellows, 2019).

Effect	Model 1	Model 2
Edges	-2.8248 (0.1103)***	-2.7212 (0.1276)***
Reciprocity	-0.4218 (0.6775)	-0.6095 (0.7846)
GW in-degree ($\alpha = 0.2$)	-0.9906 (0.1547)***	-0.9640 (0.1576)***
GW out-degree ($\alpha = 0.2$)	-0.7592 (0.1628)***	-0.7105 (0.1408)***
Two-paths	-0.0798 (0.0198)***	-0.0916 (0.0154)***
Triangles	1.1689 (0.2579)***	1.2790 (0.2533)***
Euclidean distance	-1.0682 (0.1152)***	-0.8780 (0.1040)***
Nodecov birthtime		-0.0001 (0.0000)**

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 9 shows LOLOG models for the Kaiser *C. elegans* neural network. These estimations took approximately 54 minutes and 11 minutes for Model 1 and Model 2 respectively, on a Lenovo PC with an Intel Core i5-10400 2.90 GHz CPU under Windows with cygwin (R version 4.2.1, lollog version 1.3).⁵ Model diagnostic and goodness-of-fit plots for Model 1 and Model 2 (Table 9) are shown in Figure A8 and Figure A9, respectively. These show the models fit will on the model parameters, as well as in-degree and out-degree distribution and edgewise shared partners distribution (although the fit for small values on the latter is not very good). I also estimated models without the Reciprocity (mutual) parameter, and the results were consistent, including a good fit on the distribution of reciprocal edges — note that the Reciprocity parameter in the models in Table 9 is not statistically significant.

These models show centralization on both in-degree and out-degree, and an over-representation of triangles, as well as a statistically significant negative estimate for the Euclidean distance parameter, meaning that spatially closer neurons are more likely to be connected. The centralization on degree and over-representation of triangles is consistent with the findings reported by Varshney et al. (2011), for a different version of the data, by fitting power law distributions to the degree distributions (just as in Figure 1 for the data used here), and using a conventional motif-finding method based on randomizing the network while preserving the degree of each node (Varshney et al., 2011).

The finding that triangles are over-represented in the *C. elegans* neural network was also described in the well-known “Network motifs” paper (Milo et al., 2002), and, as noted by Varshney et al. (2011), was already reported by White et al. (1986). No spatial data was used in Varshney et al. (2011), and they state that such over-representation of the triangle motif “would arise naturally if proximity was a limiting factor for connectivity, however there is no evidence for this limitation” (Varshney et al., 2011, p. 12). However, as noted by Artzy-Randrup et al. (2004), it was already shown by White et al. (1986) that neighbouring neurons are more likely than distant neurons to form a connection — and this spatial feature of clustering is not accounted for in conventional motif finding methods, as used in Milo et al. (2002); Varshney et al. (2011), where only degree distribution is preserved in the randomized networks constituting the null model. So, although spatial proximity is not necessarily a “limiting factor” for connectivity, since long distance connections do exist, and indeed in most cases can do so because the connections were formed early on in development when the neurons were spatially closer (Varier and Kaiser, 2011), it is certainly a relevant factor that should be accounted for in determining the statistical significance of the triangle motif (Artzy-Randrup et al., 2004). The LOLOG model (as does ERGM) allows for the inclusion of both the triangle structure and spatial proximity (as well as other relevant and non-independent features, including terms to model degree distribution), and as shown by the models in Table 9, both the triangle structure and spatial proximity of connected neurons are statistically significant features of this network. That is, even accounting for the spatial proximity effect (and all other effects in the model), there is significant over-representation of the triangle structure.

⁵This estimation was run on a different system from the others due to the compute cluster becoming unexpectedly, and regrettably, unavailable during the course of the work described here.

Although this model shows that increasing Euclidean distance between two neurons makes them less likely to be directly connected, this does not necessarily imply that the network has grown in such a way as to minimize the length of connections. Kaiser and Hilgetag (2006) find that neural networks (in examples including *C. elegans*) seem to minimize geodesic distance of processing paths (processing steps), rather than “wiring length”. The non-optimal wiring length arises from long-distance connections, suggesting a trade-off between wiring length and average number of processing steps (Kaiser and Hilgetag, 2006).

Model 1 (Table 9) was estimated with the node inclusion order specified in the LOLOG model as the neuronal birth times, while Model 2 does not specify the inclusion order, but instead includes the birth times as a nodal covariate. The results for both models are consistent, but Model 2 shows a negative and statistically significant estimate of the birth time covariate, indicating that older neurons tend to have more connections. This is consistent with the finding in Varier and Kaiser (2011) that early-born neurons are more highly connected than later-born neurons.

I was unable to find a converged Tapered ERGM estimate, using simple model terms such as those used for the LOLOG model (Table 9), for the Kaiser *C. elegans* neural network.

Table 10: Latent order logistic (LOLOG) model (Fellows, 2018) parameter estimates for the *Drosophila* medulla network. Estimated using the lolog package (Fellows, 2019).

Effect	Estimate	Std. error	p-value
Edges	-1.4418	0.1173	< 0.001
Reciprocity	-1.2554	0.8765	0.1521
GW in-degree ($\alpha = 0.2$)	-0.8541	0.1597	< 0.001
GW out-degree ($\alpha = 0.2$)	-0.9857	0.1320	< 0.001
Two-paths	-0.1128	0.0164	< 0.001
Triangles	1.2676	0.2523	< 0.001
Euclidean distance	-0.0012	0.0001	< 0.001

Table 11: Tapered ERGM (Fellows and Handcock, 2017; Blackburn and Handcock, 2023) parameter estimates for the *Drosophila* medulla network. Estimated using the ergm.tapered package (Handcock et al., 2022b).

Effect	Model 1
Edges	-2.579 (0.062)***
Reciprocity	1.406 (0.095)***
GW in-degree ($\alpha = 0.2$)	-3.384 (0.025)***
GW out-degree ($\alpha = 0.2$)	-3.733 (0.027)***
Two-paths	-0.000 (0.005)
Transitive triples	0.073 (0.011)***
Euclidean distance	-0.001 (0.000)***
AIC	29150.00
BIC	29229.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; \cdot $p < 0.1$.

Table 10 shows a LOLOG model for the *Drosophila* optic medulla network. This estimation took approximately 1.3 hours on a Lenovo PC with an Intel Core i5-10400 2.90 GHz CPU under Windows with cygwin (R version 4.2.1, lolog version 1.3). A Tapered ERGM model for the *Drosophila* optic medulla network is shown in Table 11. This estimation took approximately 6 minutes (on the same cluster system as the other estimations). The diagnostic and goodness-of-fit plots in Figure A10 show that the LOLOG model fits well, and the goodness-of-fit plots in Figure A11 show that the Tapered ERGM fits well, aside from the geodesic distance distribution, and some of the triad census (for example 300, three mutually connected nodes).

Both the LOLOG model (Table 10) and the Tapered ERGM model (Table 11) show centralization on both in-degree and out-degree (as we might expect from the power law and log-normal distributions fit to the degree distributions in Figure 1), and an over-representation of (transitive) triangles, as well as a negative and statistically significant estimate of the Euclidean distance parameter, indicating that spatially closer neurons are more likely to be connected.

That spatial distance between nodes is a contributor to the structure of neural networks (closer neurons are more

likely to be connected) is expected. As well as the previous findings in the context of the *C. elegans* neural network discussed above, it was found by Stillman et al. (2017) using generalized ERGM (Desmarais and Cranmer, 2012) models of the human Default Mode Network, and previously noted by Vértes et al. (2012); Roberts et al. (2016) for the human brain, and Betzel and Bassett (2017); Allard and Serrano (2020) more generally, including for *Drosophila* connectomes, and is noted in general in a review of trade-offs in brain network structure (Bullmore and Sporns, 2012). As discussed previously for the *C. elegans* neural network, that the (transitive) triangles parameter is positive and significant in a model that also includes Euclidean distance between nodes is evidence that the transitive triangle structure is over-represented, even when accounting for the spatial proximity effect.

Table 12: Summary of model fits

Network	ERGM	Tapered ERGM	LOLOG
Yeast PPI	EE only	Only fits triad census well	No
Human PPI (HIPPIE)	EE only	No	No
Alon <i>E. coli</i> regulatory	EE, statnet	Good	Good
Alon yeast regulatory	EE only	Good, except reciprocity	Good
Cook <i>C. elegans</i> connectome	No	Only fits triad census well	Poor fit on reciprocity and in-degree distribution
Kaiser <i>C. elegans</i> neural	No	No (with simple triangle terms)	Good
<i>Drosophila</i> medulla (named)	No	Poor on geodesic distance distribution and some triads	Good

“No” means a non-degenerate model was unable to be estimated. “EE” is the Equilibrium Expectation algorithm implemented in the Estimnet or EstimNetDirected software; models estimated with the EE algorithm or statnet are described in Stivala and Lomi (2021), and use the “alternating” or “geometrically weighted” model terms for triangles to avoid degeneracy. Note that the LOLOG goodness-of-fit does not include geodesic distance distribution or triad census.

Table 12 summarizes results described here for the ability to find non-degenerate models, and their goodness-of-fit, for “standard” ERGMs, Tapered ERGMs, and LOLOGs. These results show that, with the exception of the HIPPIE network (and the Yeast PPI network for LOLOG and *C. elegans* neural network for Tapered ERGM), the Tapered ERGM and LOLOG are able to estimate models for biological networks that could not be estimated using standard ERGMs with either statnet or the EstimNetDirected software. Further, the Tapered ERGM and LOLOG are able to estimate models with simple terms, such as triangle, while “alternating” (Snijders et al., 2006; Robins et al., 2007) or “geometrically weighted” (Hunter, 2007) terms are required to avoid near-degeneracy in estimating these models using standard ERGMs.

5 Conclusions and future work

I was able to estimate both Tapered ERGM and LOLOG models for biological networks for which I was unable to estimate “standard” ERGM models. Furthermore, these models can be estimated with simple model terms such as the triangle, without having to consider the problem of near-degeneracy which frequently occurs using such parameters in standard ERGMs, necessitating the use of “geometrically weighted” or “alternating” parameters instead. As discussed in Clark and Handcock (2022) for LOLOGs, and Blackburn and Handcock (2023) for Tapered ERGMs, this simplifies both the model and the modeling procedure, and allows the use of a wider set of statistics, without being constrained by problems with near-degeneracy. In the context of biological networks as considered in this work, it overcomes the limitation discussed in Stivala and Lomi (2021), that the use of more complex statistics such as “alternating k -stars” or “geometrically weighted edgewise shared partners” (GWESP) is a further step away from directly testing for over-representation of the motif. (Note that even using the triangle term is not quite the same as a motif as usually defined, since motifs are usually considered to be induced subgraphs, and ERGM configurations are usually subgraphs, not induced subgraphs, although they can also be defined to be induced subgraphs, as for example in Yaveroğlu et al. (2015)).

For the networks considered here, the LOLOG and Tapered ERGMs have consistent interpretations, and, where “standard” ERGMs were able to be estimated, their interpretations are also consistent with the LOLOG and Tapered ERGM models estimated in this work.

For networks where spatial information is available — the *C. elegans* neural network and the *Drosophila* medulla network — the ability to estimate a model with both spatial information and triangle terms overcomes the problem with standard motif significance testing methods that spatial aggregation is not accounted for in determining the over-representation of triangles (Artzy-Randrup et al., 2004). Further, using the ERGM (including Tapered ERGM) or LOLOG, we do not have to make the assumption of the geometric Chung-Lu model in Agarwala and Kenter (2023) that the “intensities” (expected degrees) of nodes and the distances between them are independent — an assumption that has biological evidence contradicting it, as noted by Agarwala and Kenter (2023).

A further advantage of the LOLOG is its ability to account for edge orderings (Fellows, 2018; Clark and Handcock, 2022). This was used in Fellows (2018) to model the order in which nodes join a social network (as they have a joining date), and although only used in the one example for which relevant information was available in Clark and Handcock (2022), it was suggested that citation networks are another natural use. In this work, I was able to use the birth times of neurons in the *C. elegans* neural network as the ordering variable in a LOLOG model, although it did not result in any clear improvement in model fit relative to using the birth time simply as a nodal covariate.

Currently, the LOLOG package has a rather limited number of model terms implemented, although there is a facility to define new model terms (Fellows, 2019). In contrast, an advantage of the Tapered ERGM is that its implementation builds on the `statnet_ergm` package, and hence can use any of the (dauntingly) many model terms available in the `ergm` package (Handcock et al., 2022a), as well as allowing user-defined terms (Hunter et al., 2013), and automatically takes advantage of computational improvements made in that software (Krivitsky et al., 2022).

For larger networks, standard estimation algorithms such as the MCMCMLE algorithm used in the `statnet_ergm` package or the Monte Carlo method of moments approach used in the LOLOG package may still not be able to estimate models in practical time. Specifically, I was only able to estimate a Tapered ERGM model for the smaller of the two PPI networks (both of which are considerably larger than the other networks considered in this work), and a LOLOG for neither of them. In contrast, ERGM models were able to be estimated for both, using the Equilibrium Expectation (EE) algorithm (Stivala and Lomi, 2021). This suggests that a fruitful line of future work could be to investigate the applicability of this algorithm to the LOLOG and the Tapered ERGM models. Another future application could be to implement the “tapering” algorithm (Fellows and Handcock, 2017; Clark and Handcock, 2022) for the autologistic actor attribute model (ALAAM), a variant of the ERGM for modeling social influence (Robins et al., 2001; Daraganova and Robins, 2013; Koskinen and Daraganova, 2022; Parker et al., 2022), which, like the ERGM, can suffer from problems of near-degeneracy, which currently can only be overcome by using “geometrically weighted” model terms (Stivala, 2023), as implemented in the ALAAMEE software (Stivala et al., 2023a).

The development of more scalable estimation algorithms (such as the EE algorithm) for LOLOG seems a particularly useful line of investigation, since very large networks can have ERGM models estimated using the EE algorithm, however such networks can be too large for the simulation-based goodness-of-fit procedure usually used for ERGMs, due to the computational difficulty of the MCMC procedure for ERGM simulation (Stivala et al., 2020), a limitation which is yet to be overcome. In contrast, a “key advantage of the LOLOG model is the ease of simulation from the model” (Clark and Handcock, 2022, p. 570); it requires only a draw from the distribution of edge orderings and sequential logistic regression on the change statistics (Fellows, 2018). This suggests that, unlike ERGMs, this simulation procedure would be practical even on very large networks, which are too large for LOLOG models to be estimated in practical time with the algorithms currently implemented.

Funding

This work was funded by the Swiss National Science Foundation (SNSF) project number 200778.

Acknowledgements

I used the high performance computing cluster at the Institute of Computing, Università della Svizzera italiana, for data processing and computations.

Data availability statement

Source code, scripts, configuration files, and datasets are freely available from https://github.com/stivalaa/bionetworks_estimations.

References

- D. Abul-Fottouh, M. Y. Song, and A. Gruzd. Examining algorithmic biases in YouTube’s recommendations of vaccine videos. *International Journal of Medical Informatics*, 140:104175, 2020. doi: <https://doi.org/10.1016/j.ijmedinf.2020.104175>.
- S. Agarwala and F. Kenter. A geometric Chung–Lu model and the *Drosophila* medulla connectome. *Journal of Complex Networks*, 11(3):cnad010, 2023. doi: 10.1093/comnet/cnad010.
- G. Alanis-Lobato, M. A. Andrade-Navarro, and M. H. Schaefer. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414, 10 2016.
- A. Allard and M. Á. Serrano. Navigable maps of structural brain networks across species. *PLoS Computational Biology*, 16(2):e1007584, 2020.
- Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone. Comment on “network motifs: simple building blocks of complex networks” and “superfamilies of evolved and designed networks”. *Science*, 305(5687):1107–1107, 2004.
- H. Bengtsson. *R.matlab: Read and Write MAT Files and Call MATLAB from Within R*, 2018. URL <https://CRAN.R-project.org/package=R.matlab>. R package version 3.6.2.
- A. Bénichou, J.-B. Masson, and C. L. Vestergaard. Compression-based inference of network motif sets. *arXiv preprint arXiv:2311.16308*, 2023.
- R. F. Betzel and D. S. Bassett. Generative models for network neuroscience: prospects and promise. *Journal of The Royal Society Interface*, 14(136):20170623, 2017.
- B. Blackburn and M. S. Handcock. Practical network modeling via tapered exponential-family random graph models. *Journal of Computational and Graphical Statistics*, 32(2):388–401, 2023.
- A. Borisenko, M. Byshkin, and A. Lomi. A simple algorithm for scalable Monte Carlo inference. *arXiv preprint arXiv:1901.00533v4*, 2020.
- E. Bullmore and O. Sporns. The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349, 2012.
- M. Byshkin, A. Stivala, A. Mira, R. Krause, G. Robins, and A. Lomi. Auxiliary parameter MCMC for exponential random graph models. *Journal of Statistical Physics*, 165(4):740–754, 2016.
- M. Byshkin, A. Stivala, A. Mira, G. Robins, and A. Lomi. Fast maximum likelihood estimation via equilibrium expectation for large network data. *Scientific Reports*, 8:11509, 2018.
- A. Caimo and I. Gollini. Recent advances in exponential random graph modelling. In *Mathematical Proceedings of the Royal Irish Academy*, volume 123, pages 1–12. Royal Irish Academy, 2023. doi: 10.1353/mpr.2023.0000.
- D. Ceperley and M. Dewing. The penalty method for random walks with uncertain energies. *Journal of Chemical Physics*, 110(20):9812–9820, 1999.
- S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461, 2013.
- Y. Choe, B. McCormick, and W. Koh. Network connectivity analysis on the temporally augmented *C. elegans* web: A pilot study. *Soc Neurosci Abstr*, 30(921.9), 2004.
- G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1:58–71, 2019.
- D. A. Clark and M. S. Handcock. Comparing the real-world performance of exponential-family random graph models and latent order logistic models for social network analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(2):566–587, 2022.

- A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.
- S. J. Cook, T. A. Jarrell, C. A. Brittin, Y. Wang, A. E. Bloniarz, M. A. Yakovlev, K. C. Nguyen, L. T.-H. Tang, E. A. Bayer, J. S. Duerr, et al. Whole-animal connectomes of both *Caenorhabditis elegans* sexes. *Nature*, 571(7763):63–71, 2019.
- M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrels. YPDTM, PombePDTM and WormPDTM: model organism volumes of the BioKnowledgeTM Library, an integrated resource for protein information. *Nucleic Acids Research*, 29(1):75–79, 01 2001. doi: 10.1093/nar/29.1.75.
- G. Csárdi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*:1695, 2006. URL <https://igraph.org>.
- G. Daraganova and G. Robins. Autologistic actor attribute models. In D. Lusher, J. Koskinen, and G. Robins, editors, *Exponential Random Graph Models for Social Networks*, chapter 9, pages 102–114. Cambridge University Press, New York, 2013.
- J. De Las Rivas and C. Fontanillo. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, 6(6):e1000807, 2010.
- E. de Silva and M. P. Stumpf. Complex networks and simple models in biology. *Journal of The Royal Society Interface*, 2(5):419–430, 2005.
- B. A. Desmarais and S. J. Cranmer. Statistical inference for valued-edge networks: The generalized exponential random graph model. *PLoS ONE*, 7(1):e30136, 2012.
- V. Dichio and F. De Vico Fallani. The exploration-exploitation paradigm for networked biological systems. *arXiv preprint arXiv:2306.17300v1*, 2023a.
- V. Dichio and F. De Vico Fallani. Statistical models of complex brain networks: a maximum entropy approach. *Reports on Progress in Physics*, 86(10):102601, 2023b. doi: 10.1088/1361-6633/ace6bc.
- F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology*, 2, 2014. doi: 10.3389/fcell.2014.00038.
- I. Fellows and M. Handcock. Removing phase transitions from Gibbs measures. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 289–297, 20–22 Apr 2017.
- I. E. Fellows. A new generative statistical model for graphs: The latent order logistic (LOLOG) model. *arXiv preprint arXiv:1804.04583v1*, 2018.
- I. E. Fellows. *lolog: Latent Order Logistic Graph Models*, 2019. URL <https://CRAN.R-project.org/package=lolog>. R package version 1.2.
- D. Felmlee, C. McMillan, and R. Whitaker. Dyads, triads, and tetrads: a multivariate simulation approach to uncovering network motifs in social graphs. *Applied Network Science*, 6(1):63, 2021.
- J. Fodor, M. Brand, R. J. Stones, and A. M. Buckle. Intrinsic limitations in mainstream methods of identifying network motifs in biology. *BMC Bioinformatics*, 21:165, 2020.
- D. Frenkel, K. J. Schrenk, and S. Martiniani. Monte Carlo sampling for stochastic weight functions. *Proceedings of the National Academy of Sciences of the USA*, 114(27):6924–6929, 2017.
- C. J. Geyer and E. A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 54(3):657–699, 1992.

- S. Ghafouri and S. H. Khasteh. A survey on exponential random graph models: an application perspective. *PeerJ Computer Science*, 6:e269, 2020.
- F. Giacomarra, G. Bet, and A. Zocca. Generating synthetic power grids using exponential random graphs models. *arXiv preprint arXiv:2310.19662v1*, 2023.
- C. S. Gillespie. Fitting heavy tailed distributions: The powerLaw package. *Journal of Statistical Software*, 64(2), 2015.
- E. Gross, S. Petrović, and D. Stasi. Goodness of fit for log-linear ERGMs. *arXiv preprint arXiv:2104.03167v5*, 2023.
- A. Gruzd, D. Abul-Fottouh, M. Y. Song, and A. Saiphoo. From Facebook to YouTube: The potential exposure to COVID-19 anti-vaccine videos on social media. *Social Media + Society*, 9(1):20563051221150403, 2023. doi: 10.1177/20563051221150403.
- M. S. Handcock. Assessing degeneracy in statistical models of social networks. Technical Report 39, Center for Statistics and the Social Sciences, University of Washington, 2003. URL <https://csss.uw.edu/Papers/wp39.pdf>.
- M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, Morris, and Martina. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1–11, 2008. URL <http://www.jstatsoft.org/v24/i01>.
- M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, P. N. Krivitsky, S. Bender-deMoll, and M. Morris. *statnet: Software Tools for the Statistical Analysis of Network Data*. The Statnet Project (<http://www.statnet.org>), 2016. URL <http://CRAN.R-project.org/package=statnet>. R package version 2016.9.
- M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, P. N. Krivitsky, and M. Morris. *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (<http://www.statnet.org>), 2022a. URL <http://CRAN.R-project.org/package=ergm>. R package version 4.3.1.
- M. S. Handcock, P. N. Krivitsky, and I. Fellows. *ergm.tapered: Tapered Exponential-Family Models for Networks*, 2022b. URL <https://github.com/statnet/ergm.tapered>. R package version 1.1-0.
- R. M. Hummel, D. R. Hunter, and M. S. Handcock. Improving simulation-based algorithms for fitting ERGMs. *Journal of Computational and Graphical Statistics*, 21(4):920–939, 2012.
- D. R. Hunter. Curved exponential family models for social networks. *Social Networks*, 29(2):216–230, 2007.
- D. R. Hunter and M. S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006.
- D. R. Hunter, P. N. Krivitsky, and M. Schweinberger. Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4):856–882, 2012.
- D. R. Hunter, S. M. Goodreau, and M. S. Handcock. *ergm.userterms: A template package for extending statnet*. *Journal of Statistical Software*, 52(2):1–25, 2013. doi: 10.18637/jss.v052.i02.
- M. Kaiser and C. C. Hilgetag. Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS Computational Biology*, 2(7):e95, 2006.
- J. Koskinen. Exponential random graph modelling. In P. Atkinson, S. Delamont, A. Cernat, J. Sakshaug, and R. Williams, editors, *SAGE Research Methods Foundations*. SAGE, London, 2020. doi: 10.4135/9781526421036888175.
- J. Koskinen. Exponential random graph models. In J. McLevey, J. Scott, and P. J. Carrington, editors, *The Sage Handbook of Social Network Analysis*, chapter 33. Sage, second edition, 2023.
- J. Koskinen and G. Daraganova. Exponential random graph model fundamentals. In D. Lusher, J. Koskinen, and G. Robins, editors, *Exponential Random Graph Models for Social Networks*, chapter 6, pages 49–76. Cambridge University Press, New York, 2013.

- J. Koskinen and G. Daraganova. Bayesian analysis of social influence. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(4):1855–1881, 2022.
- J. H. Koskinen, G. L. Robins, P. Wang, and P. E. Pattison. Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Social Networks*, 35(4):514–527, 2013.
- P. N. Krivitsky, D. R. Hunter, M. Morris, and C. Klumb. ergm 4: Computational improvements. *arXiv preprint arXiv:2203.08198v1*, 2022.
- P. N. Krivitsky, D. R. Hunter, M. Morris, and C. Klumb. ergm 4: New features for analyzing exponential-family random graph models. *Journal of Statistical Software*, 105(6):1–44, 2023. doi: 10.18637/jss.v105.i06.
- P. Leifeld, S. J. Cranmer, and B. A. Desmarais. Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. *Journal of Statistical Software*, 83(6):1–36, 2018. doi: 10.18637/jss.v083.i06.
- M. Levy. gwdegree: Improving interpretation of geometrically-weighted degree estimates in exponential random graph models. *Journal of Open Source Software*, 1(3):36, 2016.
- M. Levy, M. Lubell, P. Leifeld, and S. Cranmer. Interpretation of gw-degree estimates in ERGMs, June 2016. URL <https://doi.org/10.6084/m9.figshare.3465020.v1>.
- D. Lusher, J. Koskinen, and G. Robins, editors. *Exponential Random Graph Models for Social Networks*. Structural Analysis in the Social Sciences. Cambridge University Press, New York, 2013.
- D. Lusher, P. Wang, J. Brennecke, J. Brailly, M. Faye, and C. Gallagher. Advances in exponential random graph models. In R. Light and J. Moody, editors, *The Oxford Handbook of Social Networks*, chapter 13, pages 234–253. Oxford University Press, 2020. doi: 10.1093/oxfordhb/9780190251765.013.18.
- P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat. Systematic topology analysis and generation using degree correlations. *ACM SIGCOMM Computer Communication Review*, 36(4):135–146, 2006.
- J. L. Martin. The structure of node and edge generation in a delusional social network. *Journal of Social Structure*, 18(1):1–22, 2017. doi: 10.21307/joss-2018-005.
- J. L. Martin. Comment on geodesic cycle length distributions in delusional and other social networks. *Journal of Social Structure*, 21(1):77–93, 2020. doi: 10.21307/joss-2020-003.
- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
- C. Obando, C. Rosso, J. Siegel, M. Corbetta, and F. De Vico Fallani. Temporal exponential random graph models of longitudinal brain networks after stroke. *Journal of The Royal Society Interface*, 19(188):20210850, 2022.
- C. Orsini, M. M. Dankulov, P. Colomer-de Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguná, G. Caldarelli, et al. Quantifying randomness in real networks. *Nature Communications*, 6:8627, 2015.
- A. Parker, F. Pallotti, and A. Lomi. New network models for the analysis of social contagion in organizations: an introduction to autologistic actor attribute models. *Organizational Research Methods*, 25(3):513–540, 2022.
- D. M. Pavlovic, P. E. Vértes, E. T. Bullmore, W. R. Schafer, and T. E. Nichols. Stochastic blockmodeling of the modules and core of the *Caenorhabditis elegans* connectome. *PLoS ONE*, 9(7):e97584, 2014.
- C. Peacock, K. Milewicz, and D. Snidal. Boilerplate in international trade agreements. *International Studies Quarterly*, 63(4):923–937, 2019. doi: 10.1093/isq/sqz069.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL <http://www.R-project.org>.
- J. A. Roberts, A. Perry, A. R. Lord, G. Roberts, P. B. Mitchell, R. E. Smith, F. Calamante, and M. Breakspear. The contribution of geometry to the human connectome. *Neuroimage*, 124:379–393, 2016.

- G. Robins, P. Pattison, and P. Elliott. Network models for social influence processes. *Psychometrika*, 66(2):161–189, 2001.
- G. Robins, T. A. B. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29:192–215, 2007.
- H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, D. Millán-Zárate, E. Díaz-Peredo, F. Sánchez-Solano, E. Pérez-Rueda, C. Bonavides-Martínez, and J. Collado-Vides. RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Research*, 29(1):72–74, 2001.
- Z. M. Saul and V. Filkov. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19):2604–2611, 2007.
- M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro. HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 7(2):e31826, 2012.
- M. H. Schaefer, T. J. Lopes, N. Mah, J. E. Shoemaker, Y. Matsuoka, J.-F. Fontaine, C. Louis-Jeune, A. J. Einfeld, G. Neumann, C. Perez-Iratxeta, et al. Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Computational Biology*, 9(1):e1002860, 2013.
- C. S. Schmid and D. R. Hunter. Improving ERGM starting values using simulated annealing. *Social Networks*, 76: 209–214, 2024. doi: 10.1016/j.socnet.2023.10.002.
- C. S. Schmid, T. H. Y. Chen, and B. A. Desmarais. Generative dynamics of Supreme Court citations: Analysis with a new statistical network model. *Political Analysis*, 30(4):515–534, 2022. doi: 10.1017/pan.2021.20.
- M. Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370, 2011. doi: 10.1198/jasa.2011.tm10747.
- M. Schweinberger, P. N. Krivitsky, C. T. Butts, and J. R. Stewart. Exponential-family models of random graphs: inference in finite, super and infinite population scenarios. *Statistical Science*, 35(4):627–662, 2020.
- T. A. Scott, N. Marantz, and N. Ulibarri. Use of boilerplate language in regulatory documents: Evidence from environmental impact statements. *Journal of Public Administration Research and Theory*, 32(3):576–590, 2021. doi: 10.1093/jopart/muab048.
- S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1):64–68, 2002.
- T. A. B. Snijders. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.
- T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, 2006.
- P. E. Stillman, J. D. Wilson, M. J. Denny, B. A. Desmarais, S. Bhamidi, S. J. Cranmer, and Z.-L. Lu. Statistical modeling of the default mode brain network reveals a segregated highway structure. *Scientific Reports*, 7(1):11694, 2017.
- A. Stivala. Reply to “Comment on geodesic cycle length distributions in delusional and other social networks”. *Journal of Social Structure*, 21(1):94–106, 2020. doi: 10.21307/joss-2020-004.
- A. Stivala. Overcoming near-degeneracy in the autologistic actor attribute model. *arXiv preprint arXiv:2309.07338v2*, 2023.
- A. Stivala and A. Lomi. Testing biological network motif significance with exponential random graph models. *Applied Network Science*, 6(1):91, 2021.
- A. Stivala and A. Lomi. A new scalable implementation of the citation exponential random graph model (cERGM) and its application to a large patent citation network. Talk presented at INSNA Sunbelt XLII conference, July 2022. URL <https://doi.org/10.5281/zenodo.7951927>.

- A. Stivala, G. Robins, and A. Lomi. Exponential random graph model parameter estimation for very large directed networks. *PLoS ONE*, 15(1):e0227804, 2020.
- A. Stivala, P. Wang, and A. Lomi. ALAAMEE. Computer software, 2023a. URL <https://github.com/stivalaa/ALAAMEE>.
- A. Stivala, P. Wang, and A. Lomi. Numbers and structural positions of women in a national director interlock network. Talk presented at INSNA Sunbelt XLIII Conference, June 2023b. URL <https://doi.org/10.5281/zenodo.8092829>.
- J. E. Sulston and H. R. Horvitz. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Developmental biology*, 56(1):110–156, 1977.
- J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thomson. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental biology*, 100(1):64–119, 1983.
- A. Suratane, M. H. Schaefer, M. J. Betts, Z. Soons, H. Mannsperger, N. Harder, M. Oswald, M. Gipp, E. Ramminger, G. Marcus, et al. Characterizing protein interactions employing a genome-wide siRNA cellular phenotyping screen. *PLoS Computational Biology*, 10(9):e1003814, 2014.
- S.-y. Takemura, A. Bharioke, Z. Lu, A. Nern, S. Vitaladevuni, P. K. Rivlin, W. T. Katz, D. J. Olbris, S. M. Plaza, P. Winston, et al. A visual motion detection circuit suggested by *Drosophila* connectomics. *Nature*, 500(7461):175–181, 2013.
- S. Varier and M. Kaiser. Neural development features: spatio-temporal development of the *Caenorhabditis elegans* neuronal network. *PLoS Computational Biology*, 7(1):e1001044, 2011.
- L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. Structural properties of the *Caenorhabditis elegans* neuronal network. *PLoS Computational Biology*, 7(2):e1001066, 2011.
- G. G. Vega Yon, A. Slaughter, and K. de la Haye. Exponential random graph models for little networks. *Social Networks*, 64:225–238, 2021. doi: <https://doi.org/10.1016/j.socnet.2020.07.005>.
- P. E. Vértes, A. F. Alexander-Bloch, N. Gogtay, J. N. Giedd, J. L. Rapoport, and E. T. Bullmore. Simple models of human brain functional networks. *Proceedings of the National Academy of Sciences of the USA*, 109(15):5868–5873, 2012. doi: [10.1073/pnas.1111738109](https://doi.org/10.1073/pnas.1111738109).
- J. T. Vogelstein, E. Perlman, B. Falk, A. Baden, W. Gray Roncal, V. Chandrashekhar, F. Collman, S. Seshamani, J. L. Patsolic, K. Lillaney, et al. A community-developed open-source computational ecosystem for big neuro data. *Nature methods*, 15(11):846–847, 2018.
- C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.
- P. Wang, G. Robins, and P. Pattison. *PNet: A program for the simulation and estimation of exponential random graph models*. Melbourne School of Psychological Sciences, The University of Melbourne, 2009. URL <http://www.melnet.org.au/s/PNetManual.pdf>.
- P. Wang, G. Robins, P. Pattison, and J. Koskinen. *MPNet: Program for the simulation and estimation of (p^*) exponential random graph models for multilevel networks*. Melbourne School of Psychological Sciences, The University of Melbourne, 2014. URL <http://www.melnet.org.au/s/MPNetManual.pdf>.
- P. Wang, A. Stivala, G. Robins, P. Pattison, J. Koskinen, and A. Lomi. *PNet: Program for the simulation and estimation of (p^*) exponential random graph models for multilevel networks*, 2022. URL <http://www.melnet.org.au/s/MPNetManual2022.pdf>.
- J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 314(1165):1–340, 1986.

O. N. Yaveroglu, S. M. Fitzhugh, M. Kurant, A. Markopoulou, C. T. Butts, and N. Przulj. `ergm.graphlets`: A package for ERG modeling based on graphlet statistics. *Journal of Statistical Software*, 65(12):1–29, 2015. URL <https://www.jstatsoft.org/v065/i12>.

Appendix A Supplementary figures

Goodness-of-fit diagnostics

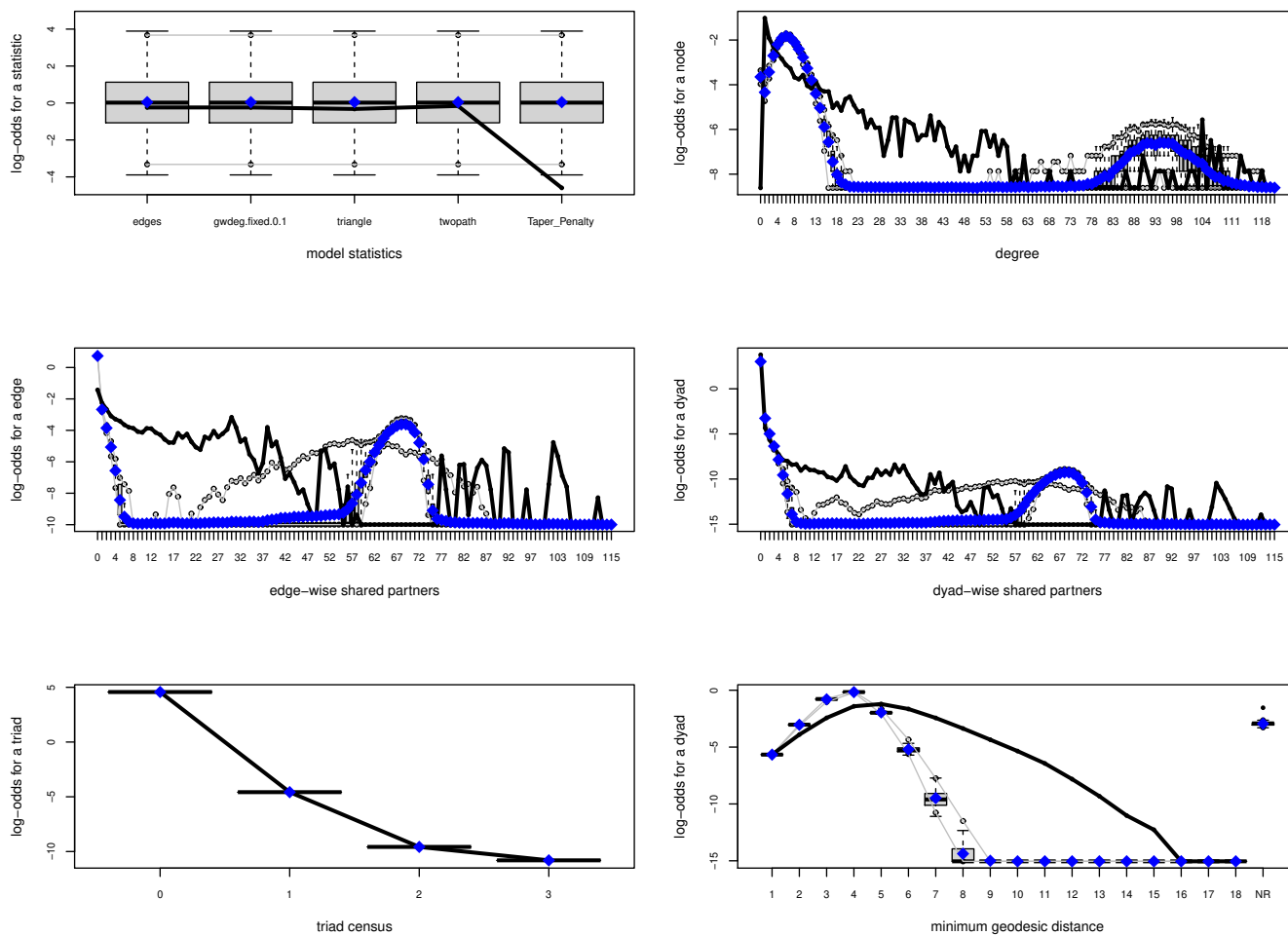
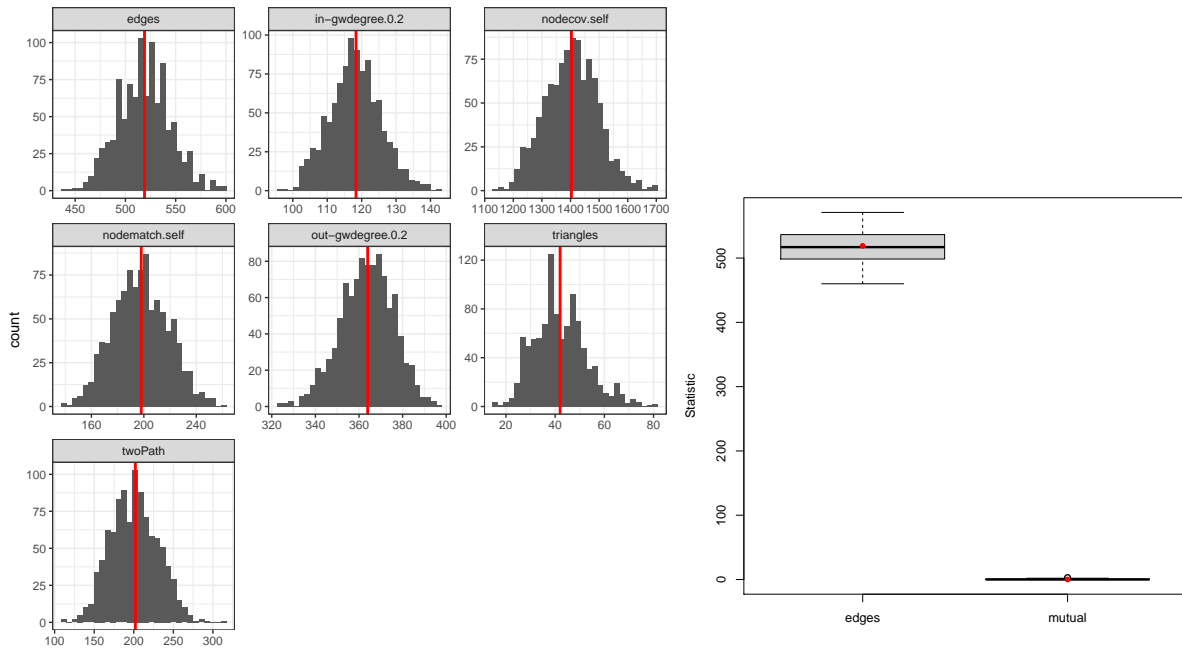
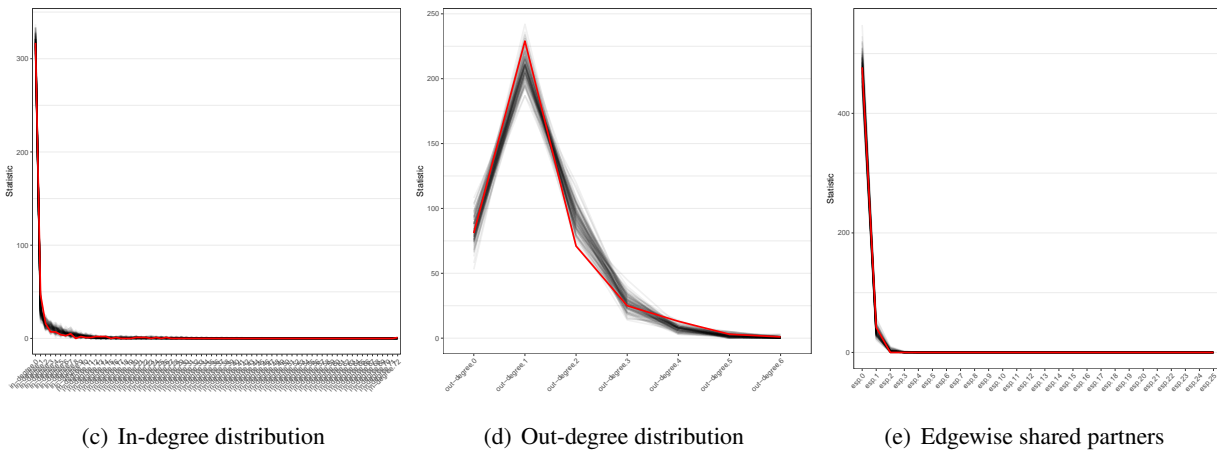


Figure A1: Goodness-of-fit plots for the tapered ERGM model of the yeast PPI network (Table 2).



(a) Model diagnostic plots

(b) Edges



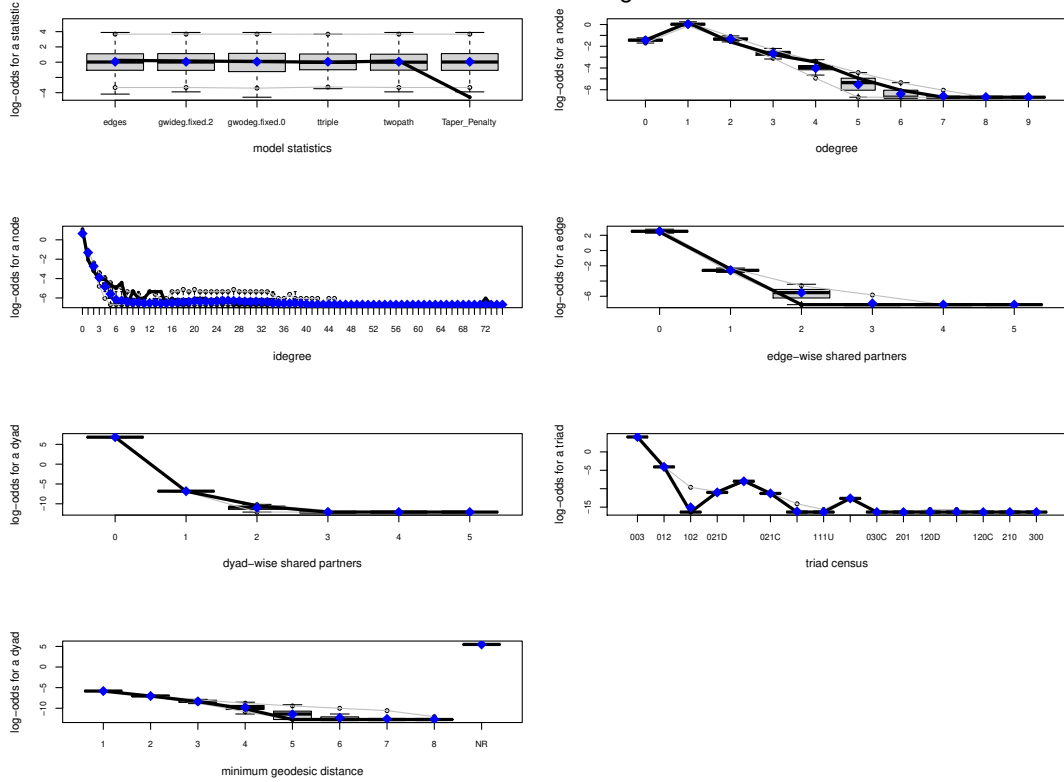
(c) In-degree distribution

(d) Out-degree distribution

(e) Edgewise shared partners

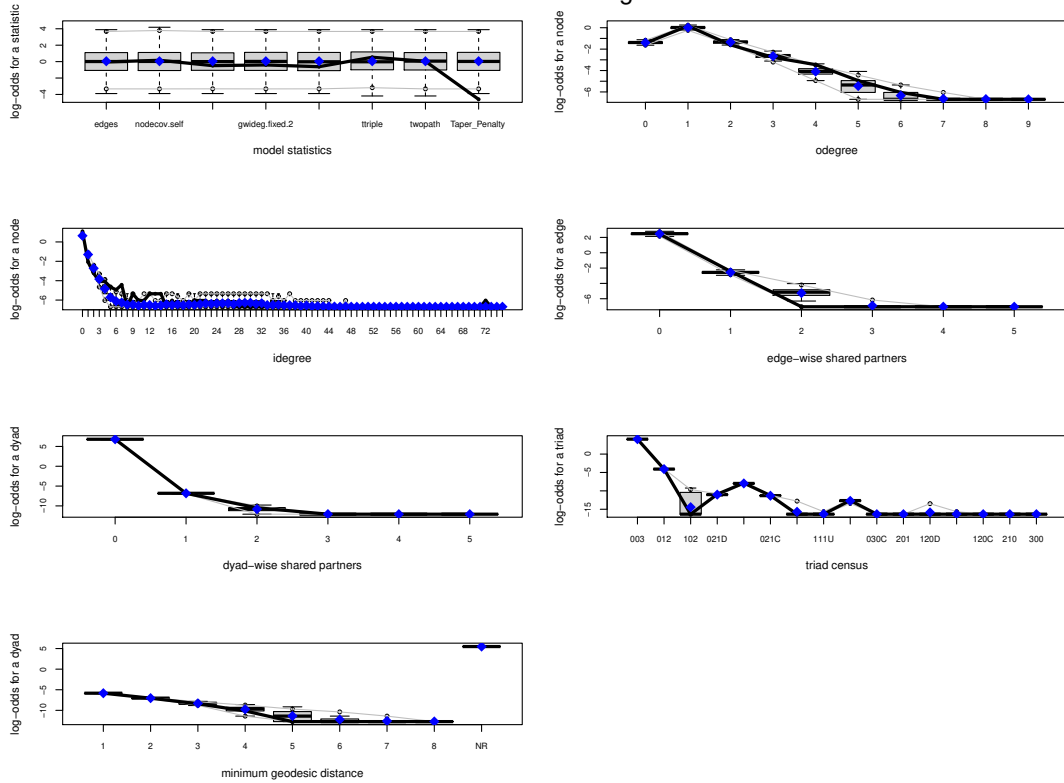
Figure A2: Model diagnostic and goodness-of-fit plots for the Alon *E. coli* regulatory network LOLOG model, Table 3.

Goodness-of-fit diagnostics



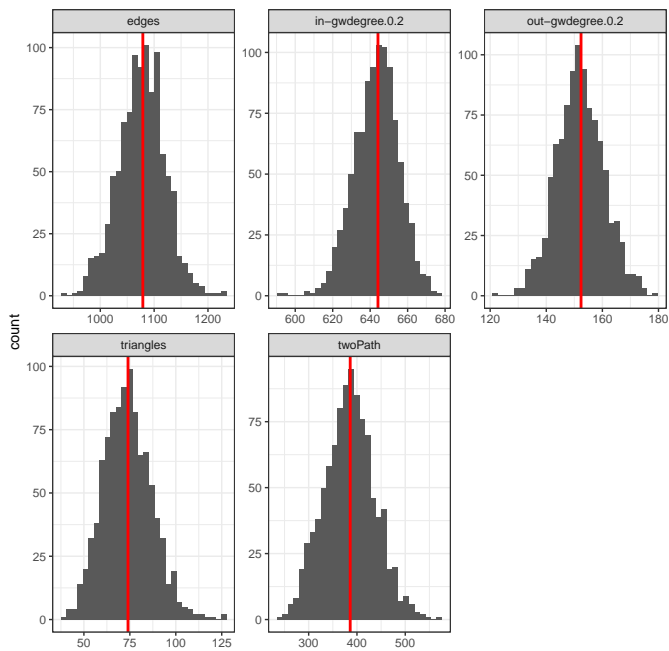
(a) Model 1

Goodness-of-fit diagnostics

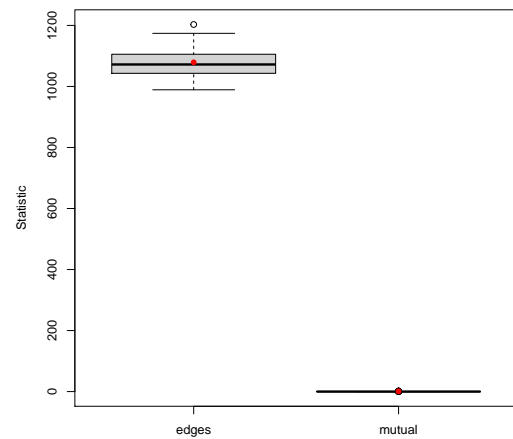


(b) Model 2

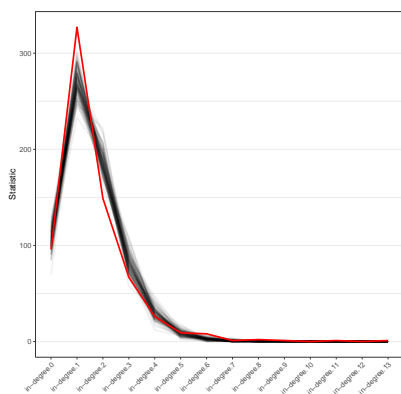
Figure A3: Goodness-of-fit plots for the tapered ERGM models of the Alon *E. coli* regulatory network (Table 4).



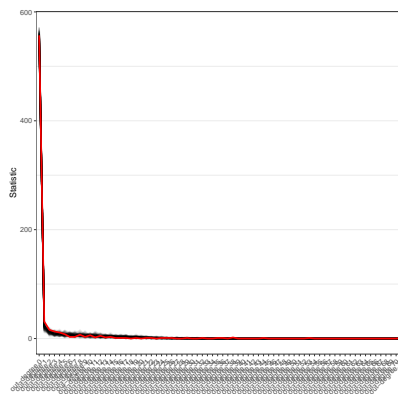
(a) Model diagnostic plots



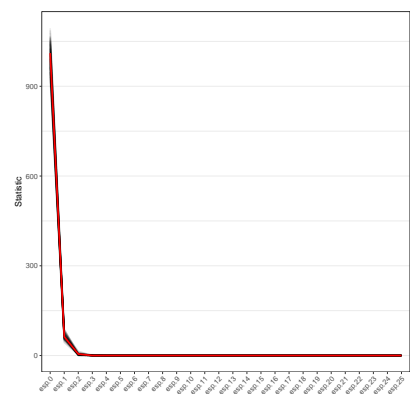
(b) Edges



(c) In-degree distribution



(d) Out-degree distribution



(e) Edgewise shared partners

Figure A4: Model diagnostic and goodness-of-fit plots for the Alon yeast network LOLOG model, Table 5.

Goodness-of-fit diagnostics

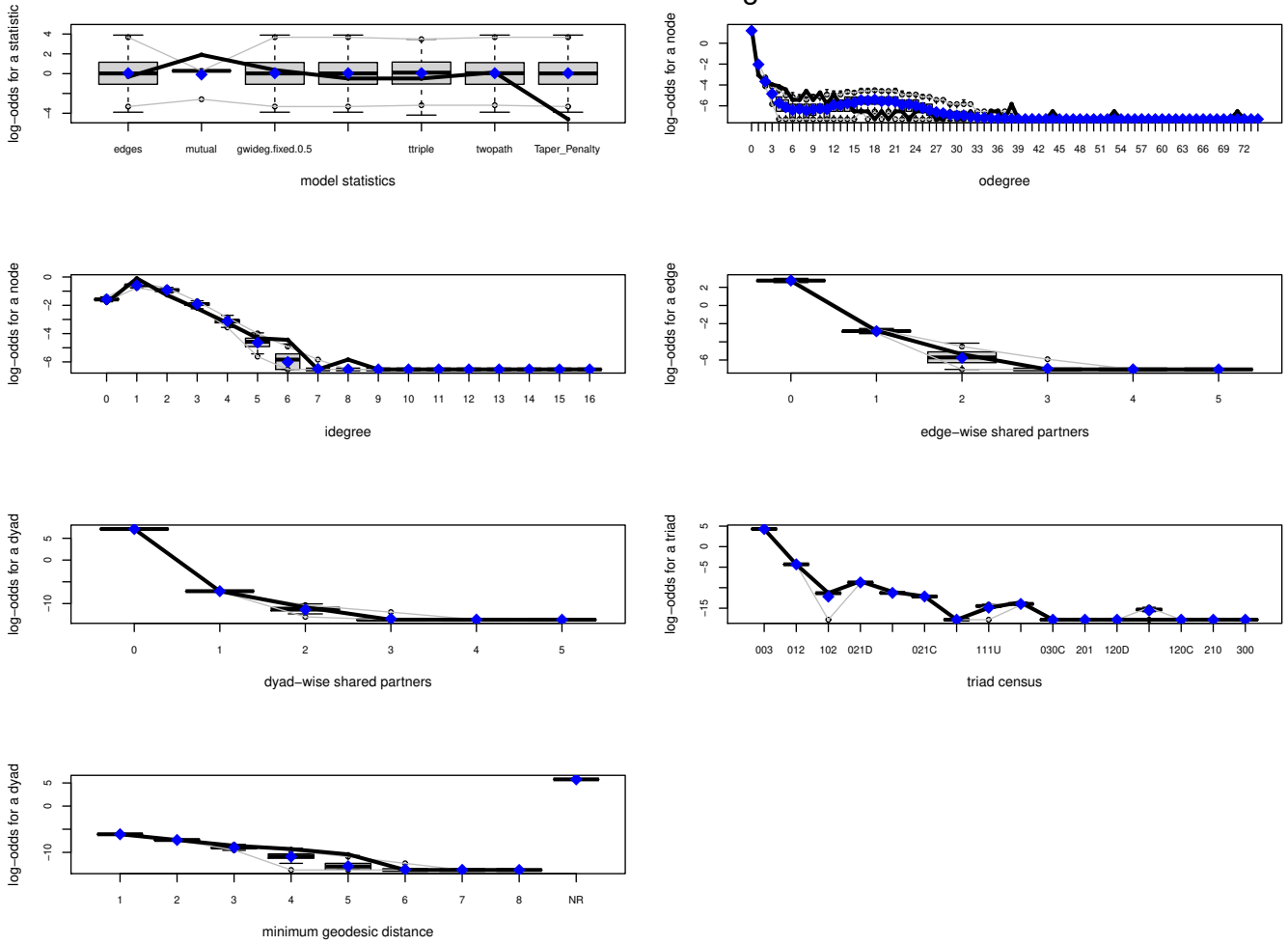
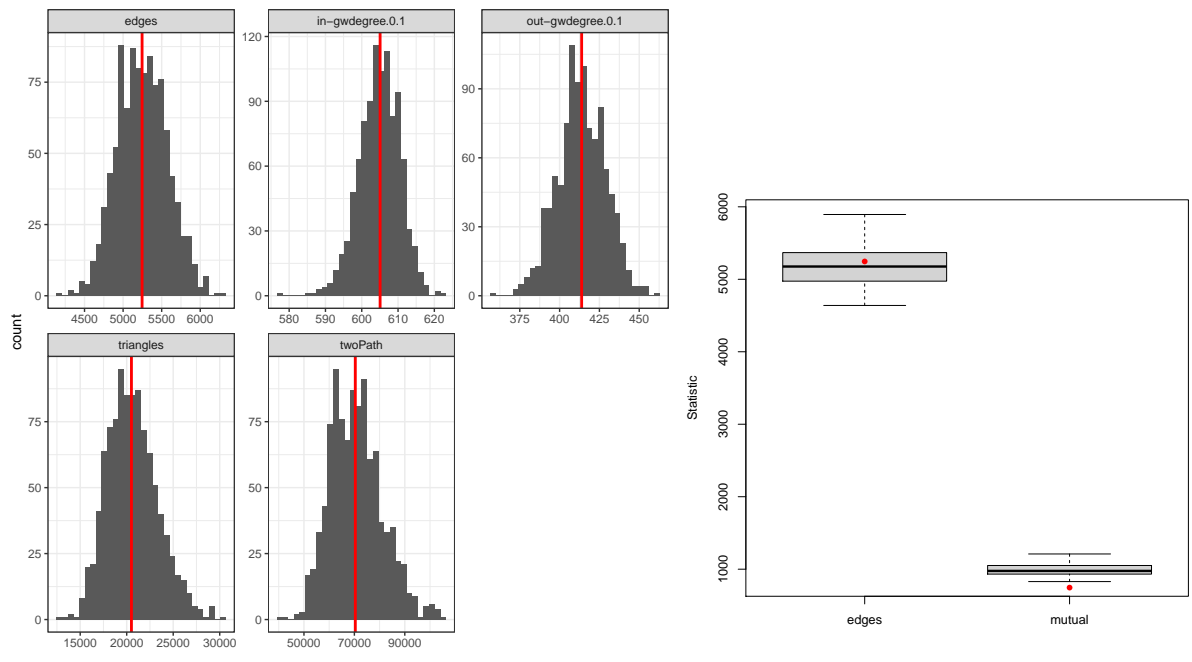
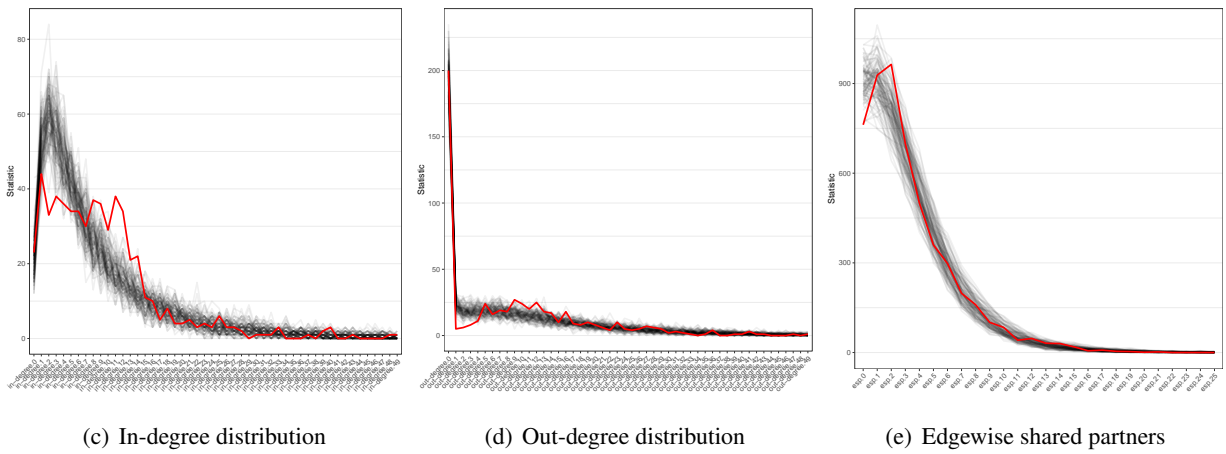


Figure A5: Goodness-of-fit plots for the tapered ERGM model of the Alon yeast regulatory network (Table 6).



(a) Model diagnostic plots

(b) Edges



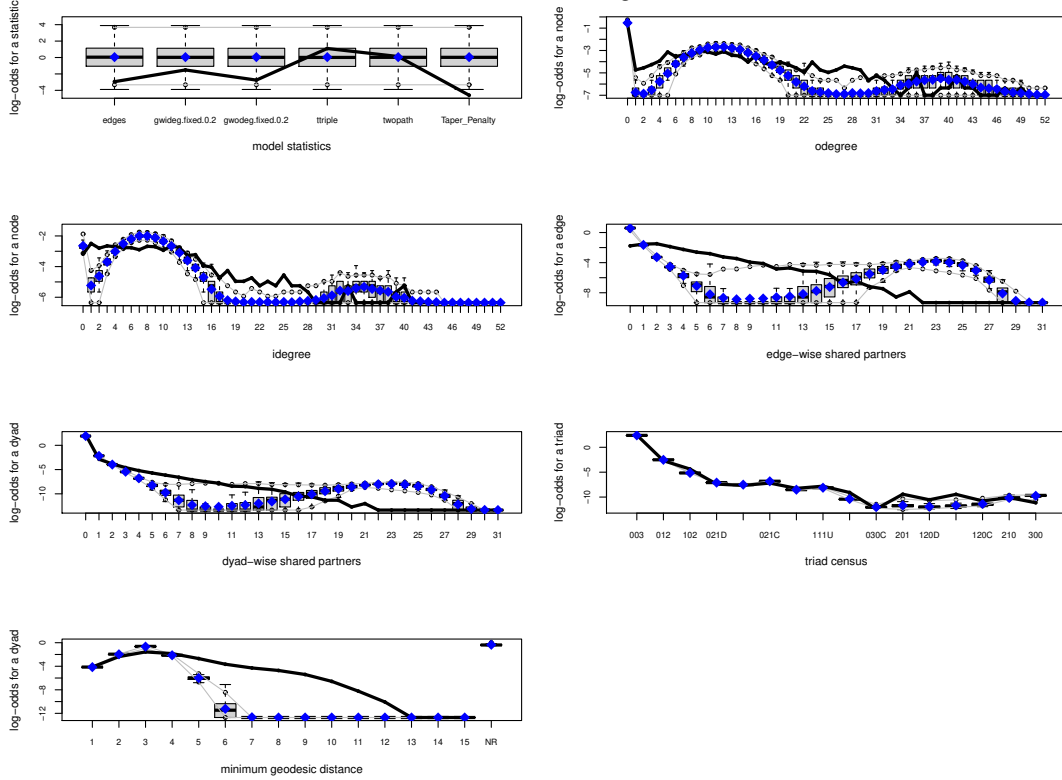
(c) In-degree distribution

(d) Out-degree distribution

(e) Edgewise shared partners

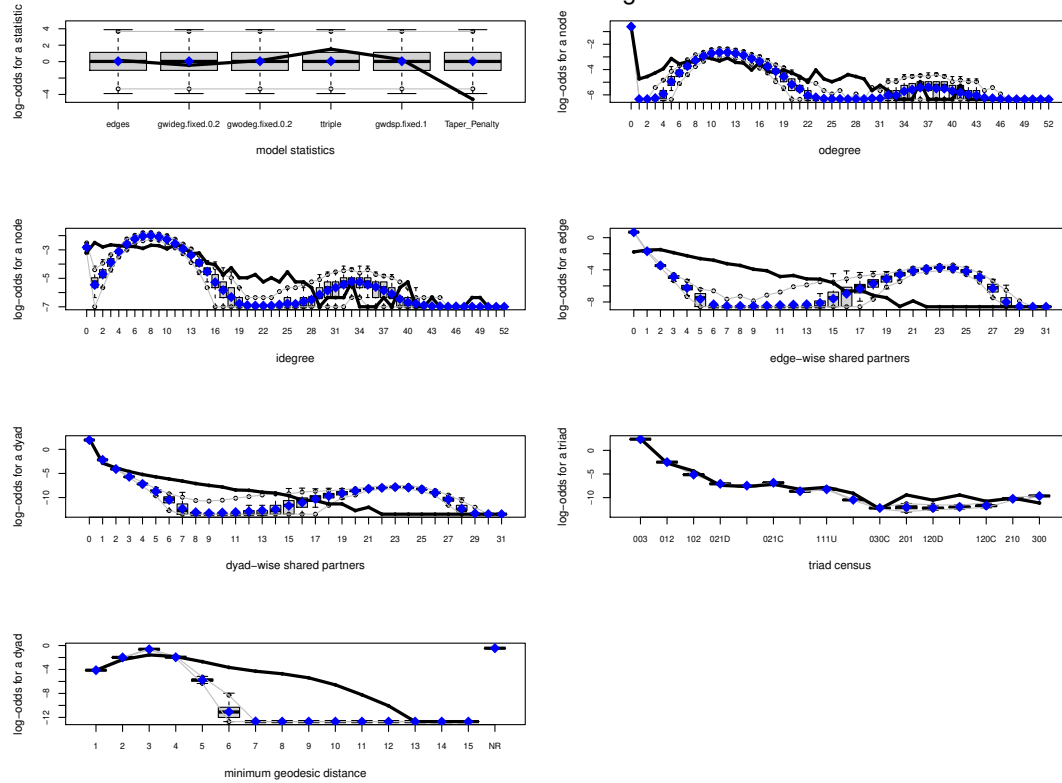
Figure A6: Model diagnostic and goodness-of-fit plots for the Cook *C. elegans* connectome LOLOG model, Table 7.

Goodness-of-fit diagnostics



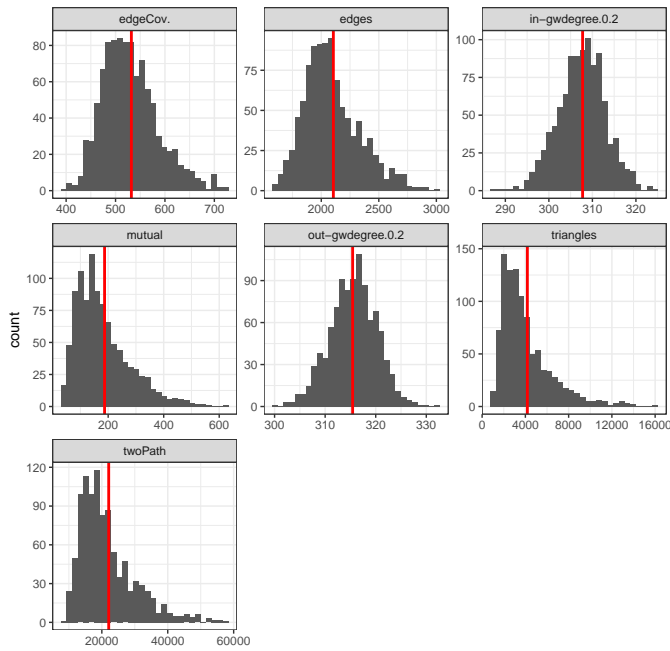
(a) Model 1

Goodness-of-fit diagnostics

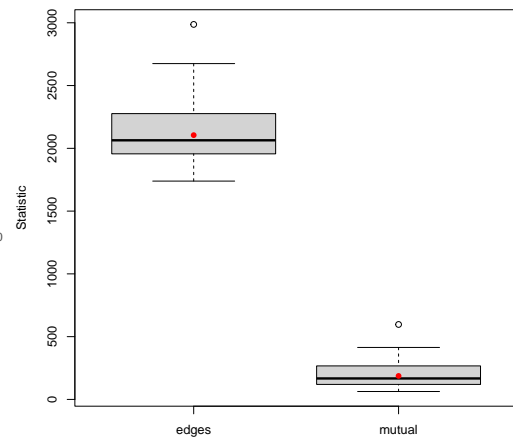


(b) Model 2

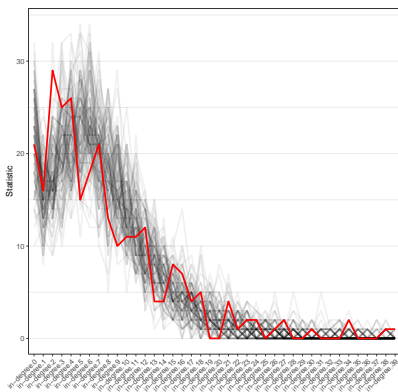
Figure A7: Goodness-of-fit plots for the tapered ERGM models of the Cook *C. elegans* connectome (Table 8).



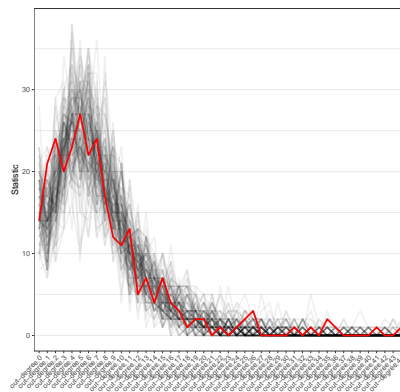
(a) Model diagnostic plots



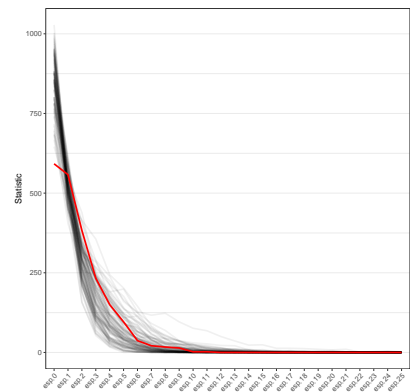
(b) Edges



(c) In-degree distribution

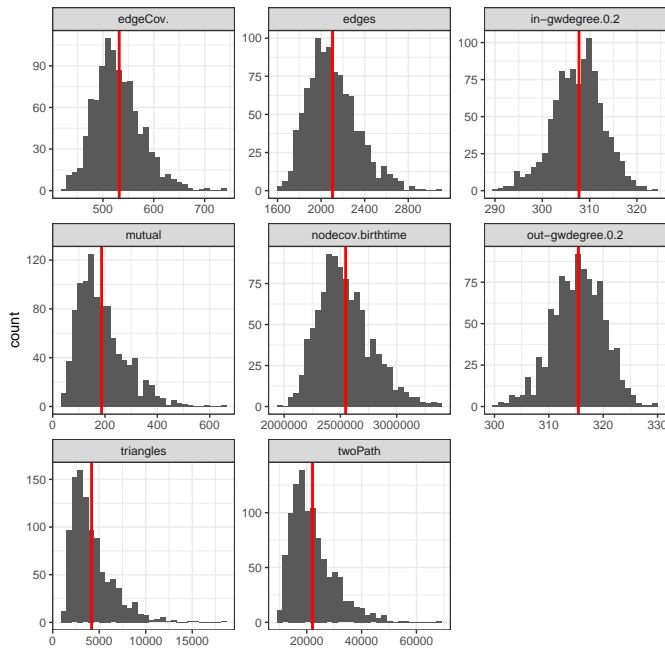


(d) Out-degree distribution

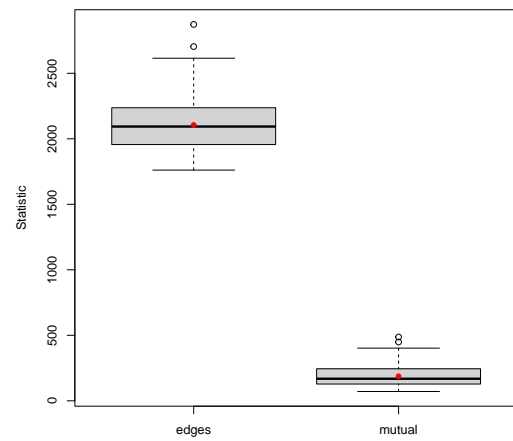


(e) Edgewise shared partners

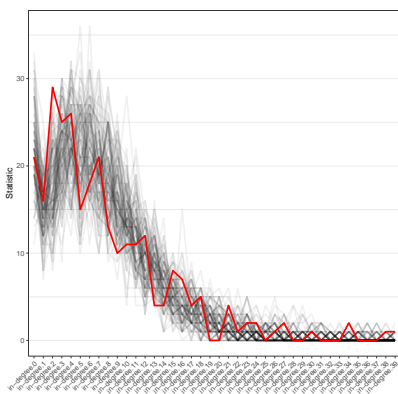
Figure A8: Model diagnostic and goodness-of-fit plots for the Kaiser *C. elegans* neural network LOLOG Model 1 (Table 9).



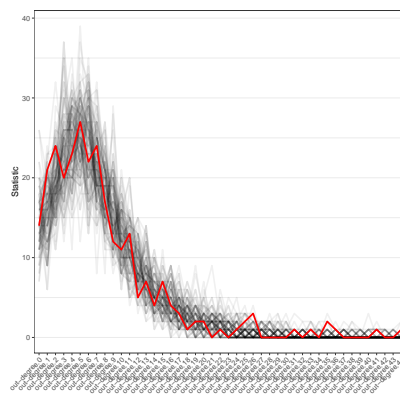
(a) Model diagnostic plots



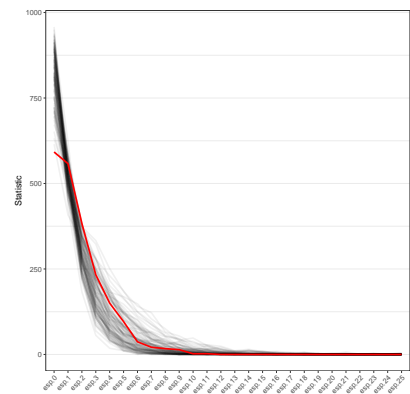
(b) Edges



(c) In-degree distribution

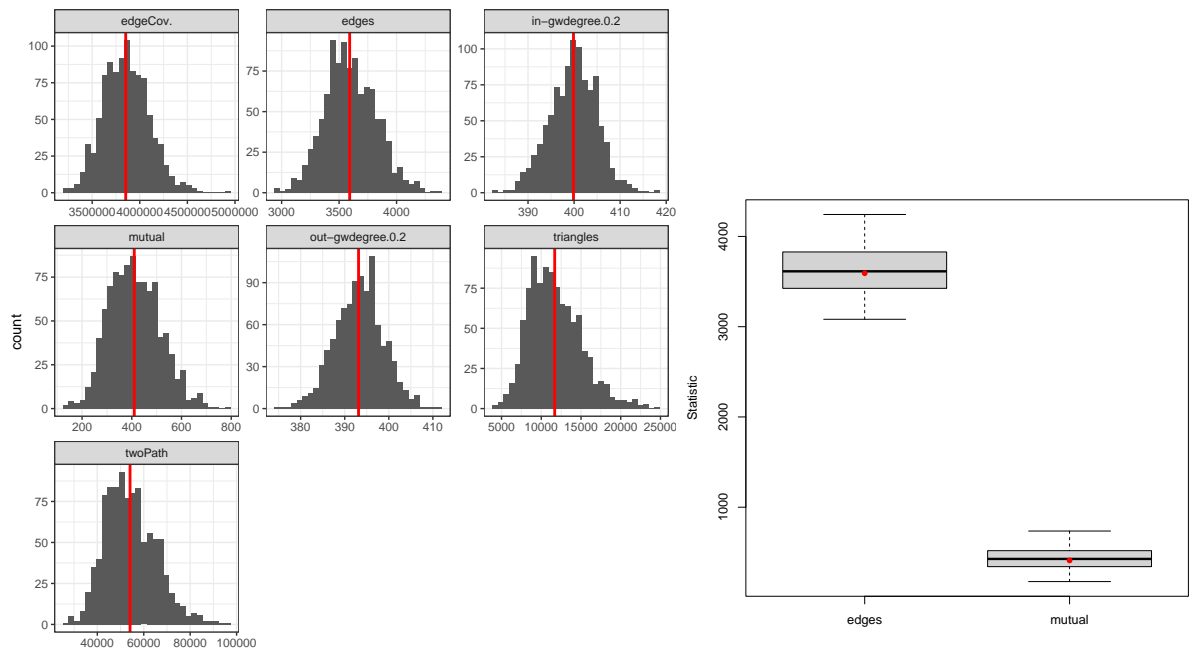


(d) Out-degree distribution



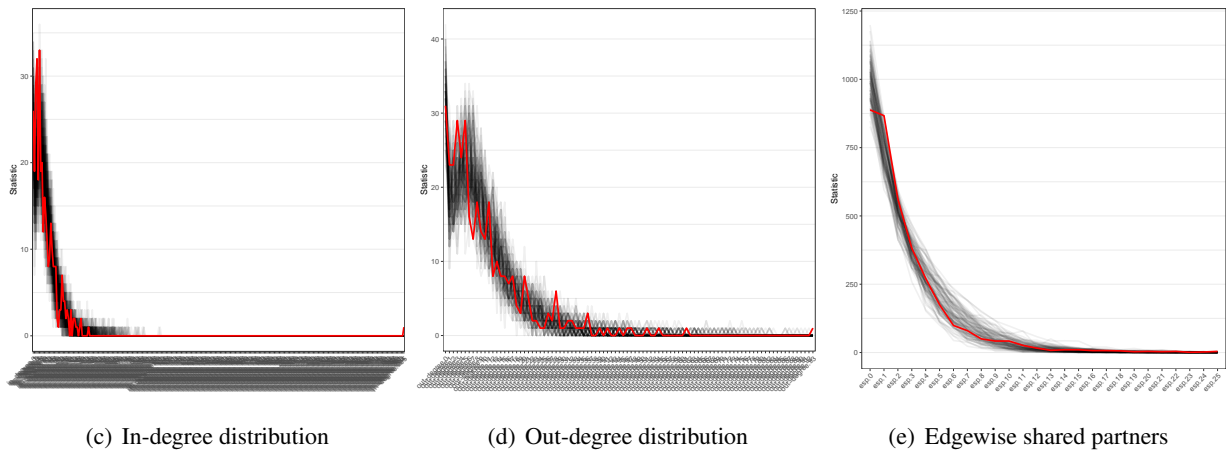
(e) Edgewise shared partners

Figure A9: Model diagnostic and goodness-of-fit plots for the Kaiser *C. elegans* neural network LOLOG Model 2 (Table 9).



(a) Model diagnostic plots

(b) Edges



(c) In-degree distribution

(d) Out-degree distribution

(e) Edgewise shared partners

Figure A10: Model diagnostic and goodness-of-fit plots for the *Drosophila* medulla network LOLOG model, Table 10.

Goodness-of-fit diagnostics

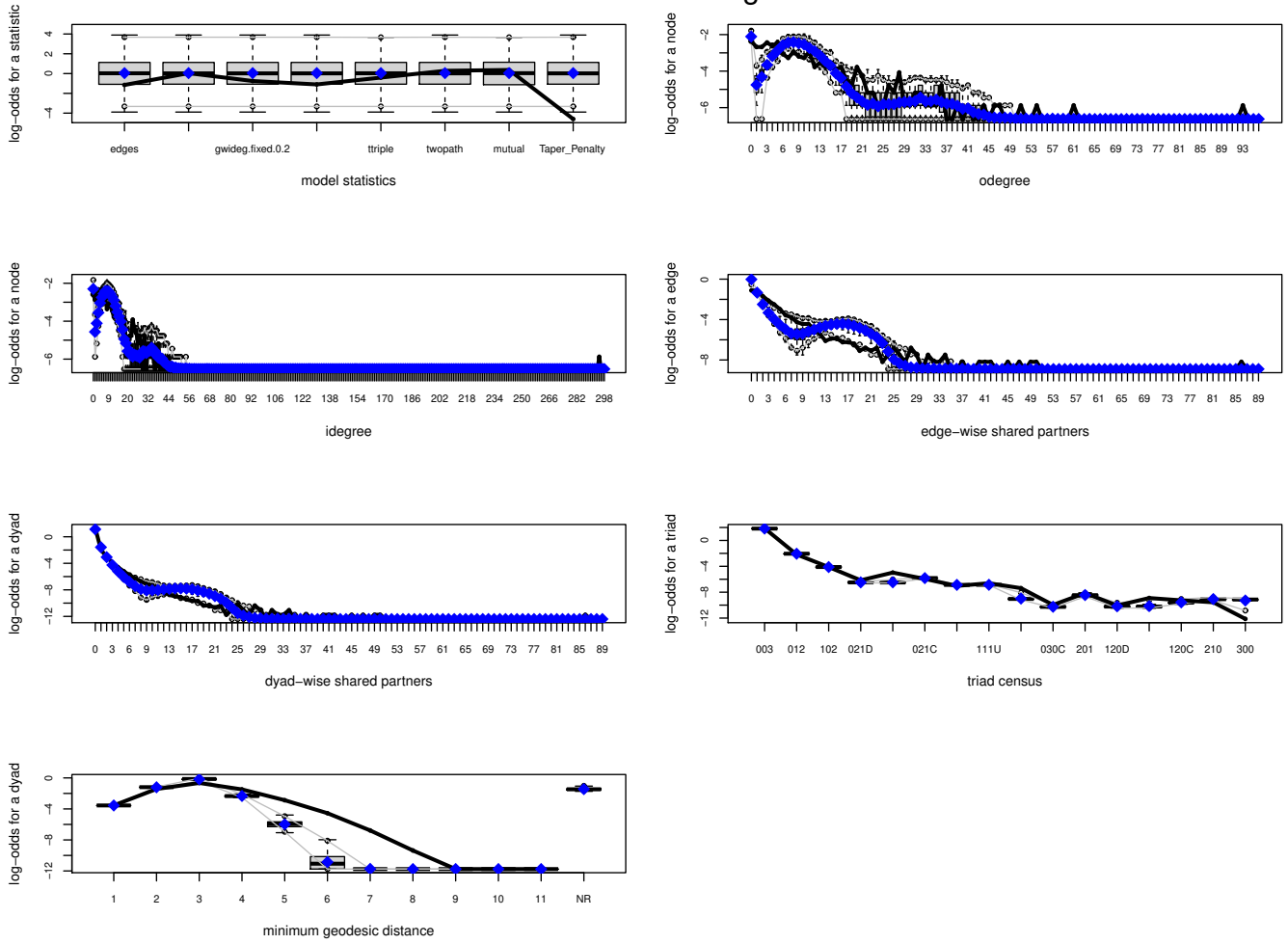


Figure A11: Goodness-of-fit plots for the tapered ERGM model of the *Drosophila* medulla network (Table 11).