Exponential-Family Random Graph Models for Multi-Layer Networks (Supplement)

Pavel N. Krivitsky* Laura M. Koehly† Christopher Steven Marcum‡

A Diagnostics for Lazega's Lawyers models

To assess goodness-of-fit for the Lazega example, 10,000 draws were simulated from the fitted Model 4 and evaluated the extent to which observed sufficient statistics were covered by the distribution of those same statistics under simulation. Specifically, the predictive distribution of the sufficient statistics associated with terms specified in our model, within-layer in- and out-degree distributions, edgewise and non-edgewise shared partners distributions, and multiple-layer degree, shared partners, and two-path distributions were all evaluated. Figure 1 reports the box-and-whisker plots of the distribution of 10,000 simulated sufficient statistics under the fitted model with the observed values superimposed as red dots. The values are logged for legibility. The model appears to replicate observed values very well, with all of the observed values falling nearly in the center of the predicted distributions—thus, lending evidence that suggests our model is not degenerate. Inspection of figure 2, which plots the comparison between the observed and simulated cross-layer in-bound/out-bound edgewise/non-edgewise shared partners distributions reveals that the model does reasonably well to capture the observed shape of the out-bound and in-bound edgewise distributions (two leftmost panels)—though the model does appear to miss the dip at the 4-count of each effect. The model does better at capturing the shape—but not the scale—of the respective nonedgewise shared partners distributions (two rightmost panels). The model does very well at capturing its lower order component two-path effects per figure 3. Notwithstanding the fact that we estimate effects for only the distinct, that is non-reciprocal, two-paths following the receipt of a coworker tie followed by sending an advice tie (leftmost most plot), the model does reasonably well covering the observed values of five out of six of the distinct cross-layer two-paths (the exception is the receiving advice to sending coworker ties two-path). This suggests that the lower order cross-layer two-path effects—which are marginal to the cross-layer edgewise shared partners distribution—are well-approximated by the joint effect of the single cross-layer two-path and cross-layer GWESP terms in Model 4. Finally, the model does a moderate job capturing the shape and scale, albeit not the exact observed values, of the multi-layer in-degree distribution and has a difficult time recovering the very

*Senior Lecturer in Statistics, The University of New South Wales, p.krivitsky@unsw.edu.au
†Senior Investigator, National Institutes of Health
‡Staff Scientist – Methodologist, National Institutes of Health

bumpy multi-layer out-degree distribution, both depicted in figure 4. Model convergence and within-layer effects were also evaluated using the same simulation procedure and with similar results—those results can be found in the supplement.

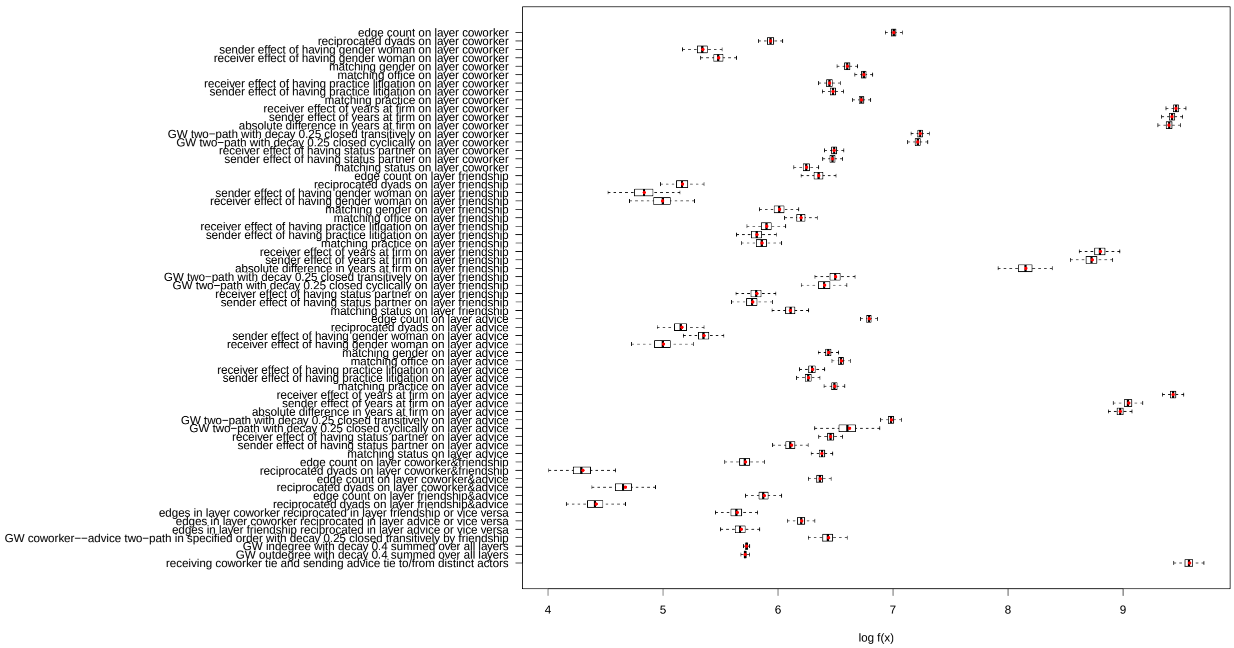

Figure 1: Lazega Data ERGM Model Box-and-Whisker plots of predicted distributions of ERGM sufficient statistics of model terms the fitted ERGM model of the Lazega Multi-layer network (n=10,000 simulations). Observed values are indicated in red.

B Simulation of Computational Time for Multi-Layer ERGMs

In general, ERGMs—even for traditional single layer networks—are notoriously inefficient on complex parameterisations and large networks with respect to convergence time. Recently, estimation techniques have been improved using Markov chain Monte Carlo to efficiently sample parameter space as well as new developments in heuristic approaches. Both the latest releases of the statnet packages for R implement these developments in the default specification of the ERGM estimation functions that we employ here. It is useful, however, to evaluate how well the state-of-the-art estimation techniques for ERGMs fair under our new multi-layer extensions to ERGM.

There are many aspects of computational complexity and runtime that could be assessed in this modeling framework. Here, we evaluate expected computational time for a few different scenarios with multi-layer cases. First, we present simulated data with 50 nodes and 500 edges (i.e., density is roughly 20%) and vary the number of layers from 2 to 10 (with constant density). To these simulated networks we fit five simple yet representative models 100 times

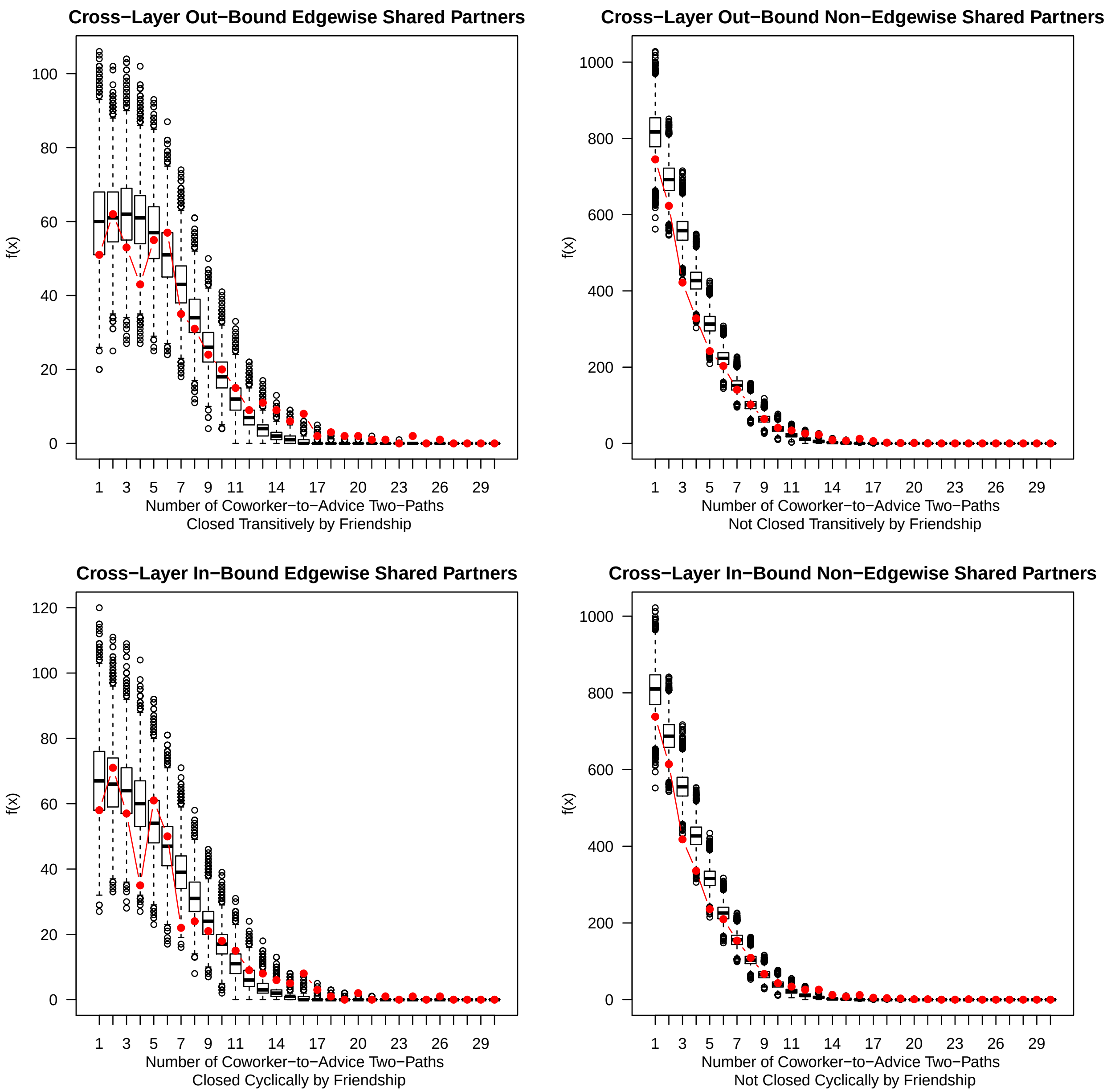

Figure 2: Lazega Data ERGM Model Box-and-Whisker plots of predicted distributions of various cross-layer shared partner distributions (n=10,000 simulations). Each two-path in each of these out-bound and in-bound shared partner distributions follows either coworker to advice layers or vice versa. Only the distribution depicted in the top left panel, out-bound two-paths, is modeled directly by a cross-layer GWESP term in the model. Observed values are indicated in red.

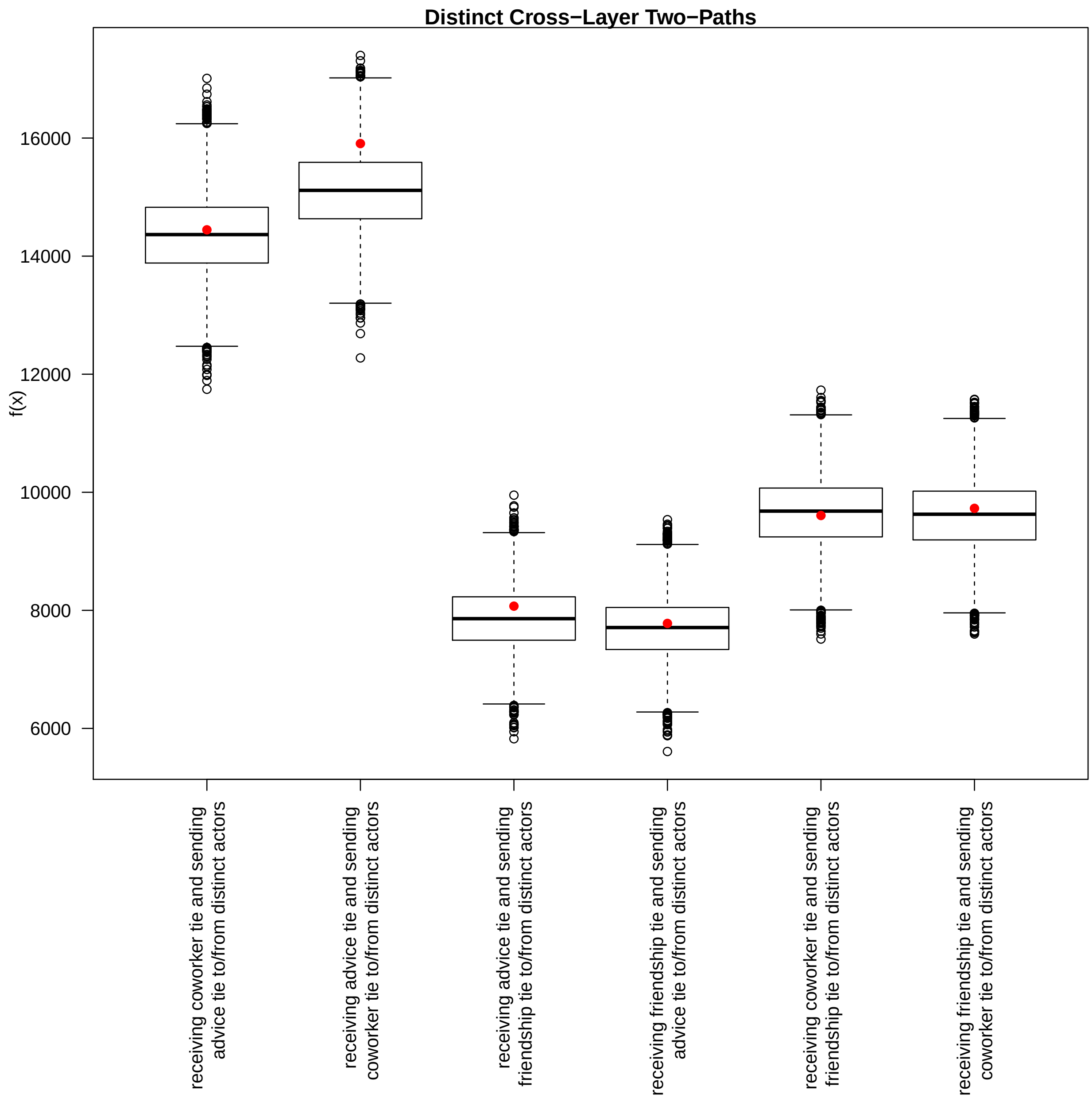

Figure 3: Lazega Data ERGM Model Box-and-Whisker plots of the predicted distributions of all six distinct cross-layer two-paths (n=10,000 simulations). Only the receiving coworker to sending advice two-path effects are modeled directly by a term in the model. Observed values are indicated in red.

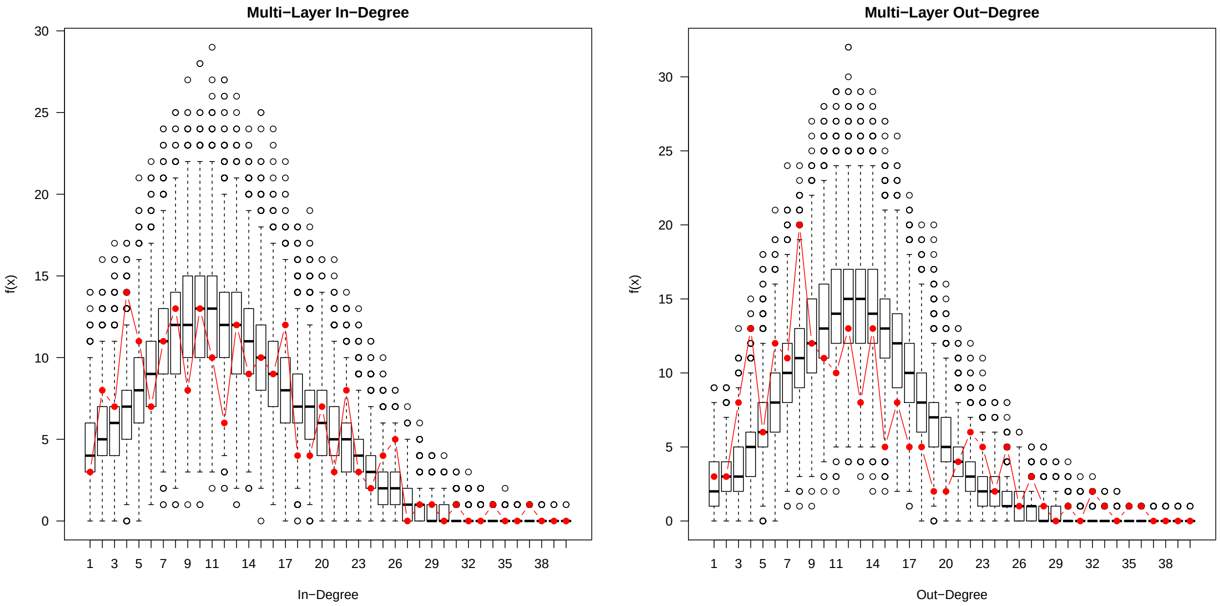

Figure 4: Lazega Data ERGM Model Box-and-Whisker plots of the predicted distribution of multi-layer in- and out- degree distributions (n=10,000 simulations). Observed values are indicated in red.

and evaluate the computational time mean and standard deviation from the sample. The five models included: 1) homogenous edges terms over all layers, 2) edges for a single logical layer conjunction of two layers, 3) edges for a single logical layer disjunction of two layers, 4) edges for a single logical layer exclusive disjunction, and 5) CMB over all layers. Second, to provide a case of expected convergence time using empirical data we refit the two most complex of the Knecht models presented in the manuscript 100 times and again evaluate the mean and standard deviation of the computational times. In both cases, we use the default ergm control specifications. Each simulation was conducted using a single node on an AMD Phenom II X2 560 processor with 6GB of RAM (mimicking a simple desktop setup) using R version 3.6.2 on the NIH's Biowulf High Performance Computation group.

Table @ref(tab:tab1) reports the computational time results from the first set of simulations. The number of simulated layers is reported in the first column and each subsequent column indexes a different model specification as detailed above. The cell values represent the mean number of seconds (with standard deviations in parentheses) that transpired between the initial model call and its completion. On average, estimation was completed quickly across all scenarios. The longest runtime was estimated at just under 2 minutes in the 10 layer CMB model. Examining the logical layer specifications along the first row reveals that one efficiency that could be made specific to multi-layer ERGMs is to ensure that only the layers under evaluation are included in the network object. Here, each logical layer specification evaluates only two layers which are quickest to estimate when only those two layers are present in the data. As such, increasing the number of layers in a multi-layer network tends to increase the computational time as evidenced in these results (the association is roughly linear but not always monotonic in every scenario). In this simulation the CMB models tended to take the longest to complete estimation.

In the second set of analyses, however, where we re-estimate the two best-fitting models of the Knecht data (pairwise logical layers with covariates and CMB with covariates) 100 times, we find a different result. Using these empirical data, we find that the logical layer specification took an order of magnitude longer to converge than the CMB specification. Specifically, the logical layer Knecht model took, on average, 5807.76(929.66) seconds to converge (or about an hour and half per fit) while the CMB model took only and average of 22.79(4.72) seconds, or about half a minute. Thus, the convergence time highly depends on the context of the specific data and the specific model being estimated. However, these results demonstrate that with typical social network data the multi-layer ERGMs can be estimated in reasonable runtimes.

Table 1: Means and standard deviations computational time (in seconds) of various multi-layer ERGM scenarios using simulated data (n=100 simulations)

No. Layers	Edges (Ind.)	Edges (Conj.)	Edges (Disj.)	Edges (XOR)	CMB
2	0.09 (0.01)	3.24 (0.41)	4.22 (0.67)	1.95 (0.51)	2.60 (0.71)
3	0.13 (0.00)	5.40 (0.72)	5.97 (0.89)	2.53 (0.69)	4.80 (1.55)
4	0.19 (0.00)	7.22 (0.85)	8.81 (1.54)	3.28 (0.80)	10.11 (3.43)
5	0.26 (0.02)	9.38 (1.19)	11.78 (1.55)	3.78 (0.96)	19.12 (4.74)
6	0.34 (0.03)	12.17 (1.55)	14.51 (2.17)	4.81 (1.49)	28.36 (7.16)
7	0.42 (0.03)	8.03 (1.15)	11.71 (2.31)	4.77 (1.19)	42.34 (9.90)
8	0.53 (0.04)	10.71 (1.73)	15.80 (2.51)	5.62 (1.59)	58.78 (16.13)
9	0.64 (0.03)	15.55 (1.76)	18.73 (2.98)	6.68 (2.20)	81.84 (23.15)
10	0.84 (0.06)	11.26 (1.89)	16.47 (2.90)	6.61 (1.92)	116.10 (28.83)