Inferential Approaches for Network Analysis: AMEN for Latent Factor Models

Shahryar Minhas[†], Peter D. Hoff[‡], & Michael D. Ward[⌐]

[†] Department of Political Science, Michigan State University
[‡] Department of Statistical Sciences, Duke University
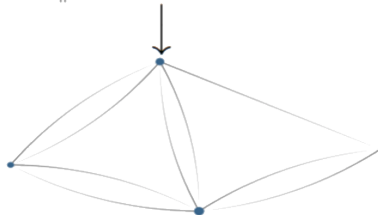[⌐] Department of Political Science, Duke University

Relational data consists of

- a set of units or nodes
- a set of measurements, y_{ij}, specific to pairs of nodes (i, j)

Sender	Receiver	Event
i	j	y_{ij}
⋮	k	y_{ik}
	l	y_{il}
j	i	y_{ji}
⋮	k	y_{jk}
	l	y_{jl}
k	i	y_{ki}
⋮	j	y_{kj}
	l	y_{kl}
l	i	y_{li}
⋮	j	y_{lj}
	k	y_{lk}

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

Relational data assumptions

GLM: $y_{ij} \sim \beta^T X_{ij} + e_{ij}$

Networks typically show evidence against independence of $e_{ij} : i \neq j$

Not accounting for dependence can lead to:

- biased effects estimation
- uncalibrated confidence intervals
- poor predictive performance
- inaccurate description of network phenomena

We've been hearing this concern for decades now:

Thompson & Walker (1982)	Beck et al. (1998)	Snijders (2011)
Frank & Strauss (1986)	Signorino (1999)	Erikson et al. (2014)
Kenny (1996)	Li & Loken (2002)	Aronow et al. (2015)
Krackhardt (1998)	Hoff & Ward (2004)	Athey et al. (2016)

Values across a row, say $\{y_{ij}, y_{ik}, y_{il}\}$, may be more similar to each other than other values in the adjacency matrix because each of these values has a common sender i

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

Values across a column, say $\{y_{ji}, y_{ki}, y_{li}\}$, may be more similar to each other than other values in the adjacency matrix because each of these values has a common receiver i

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

Actors who are more likely to send ties in a network may also be more likely to receive them

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

Values of y_{ij} and y_{ji} may be statistically dependent

	i	j	k	l
i	NA	y_{ij}	y_{ik}	y_{il}
j	y_{ji}	NA	y_{jk}	y_{jl}
k	y_{ki}	y_{kj}	NA	y_{kl}
l	y_{li}	y_{lj}	y_{lk}	NA

We use this model to form the additive effects portion of AME

$$y_{ij} = \mu + e_{ij}$$

$$e_{ij} = a_i + b_j + \epsilon_{ij}$$

$$\{(a_1, b_1), \dots, (a_n, b_n)\} \sim N(0, \Sigma_{ab})$$

$$\{(\epsilon_{ij}, \epsilon_{ji}) : i \neq j\} \sim N(0, \Sigma_\epsilon), \text{ where}$$

$$\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \quad \Sigma_\epsilon = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- μ baseline measure of network activity
- e_{ij} residual variation that we will use the SRM to decompose

Social Relations Model (The "A" in AME)

$$y_{ij} = \mu + e_{ij}$$

$$e_{ij} = a_i + b_j + \epsilon_{ij}$$

$$\{(a_1, b_1), \ldots, (a_n, b_n)\} \sim N(0, \Sigma_{ab})$$

$$\{(\epsilon_{ij}, \epsilon_{ji}) : \ i \neq j\} \sim N(0, \Sigma_\epsilon), \ \text{where}$$

$$\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \quad \Sigma_\epsilon = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- row/sender effect (a_i) & column/receiver effect (b_j)

- Modeled jointly to account for correlation in how active an actor is in sending and receiving ties

$$y_{ij} = \mu + e_{ij}$$

$$e_{ij} = a_i + b_j + \epsilon_{ij}$$

$$\{(a_1, b_1), \ldots, (a_n, b_n)\} \sim N(0, \Sigma_{ab})$$

$$\{(\epsilon_{ij}, \epsilon_{ji}) : i \neq j\} \sim N(0, \Sigma_\epsilon), \text{ where}$$

$$\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \quad \Sigma_\epsilon = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- σ_a^2 and σ_b^2 capture heterogeneity in the row and column means

- σ_{ab} describes the linear relationship between these two effects (i.e., whether actors who send [receive] a lot of ties also receive [send] a lot of ties)

$$y_{ij} = \mu + e_{ij}$$

$$e_{ij} = a_i + b_j + \epsilon_{ij}$$

$$\{(a_1, b_1), \ldots, (a_n, b_n)\} \sim N(0, \Sigma_{ab})$$

$$\{(\epsilon_{ij}, \epsilon_{ji}) : \ i \neq j\} \sim N(0, \Sigma_\epsilon), \ \text{where}$$

$$\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \qquad \Sigma_\epsilon = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- ϵ_{ij} captures the within dyad effect
- Second-order dependencies are described by σ_ϵ^2
- Reciprocity, aka within dyad correlation, represented by ρ

Third Order Dependencies

HOMOPHILY
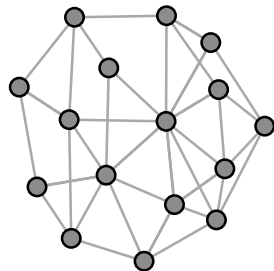
STOCHASTIC EQUIVALENCE
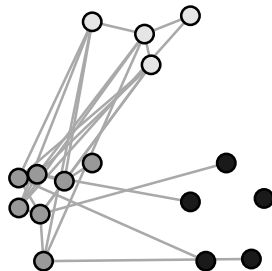
To account for these patterns we can build on what we have so far and find an expression for γ:

$$y_{ij} \approx \beta^T X_{ij} + a_i + b_j + \gamma(u_i, v_j)$$

Each node i has an unknown latent factor

$$\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^k \ \ i, j \in \{1, \ldots, n\}$$

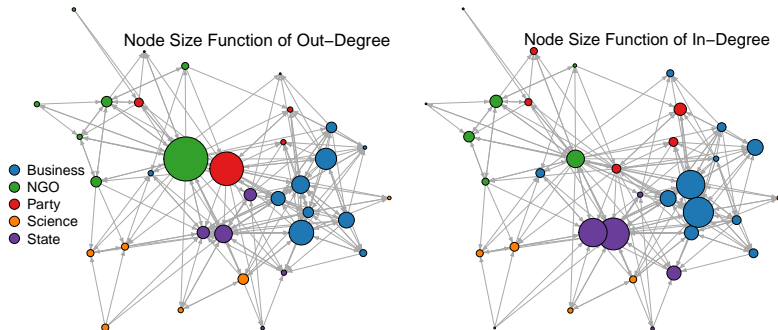The probability of a tie from i to j depends on their latent factors

$$\gamma(\mathbf{u}_i, \mathbf{v}_j) = \mathbf{u}_i^T D \mathbf{v}_j$$
$$= \sum_{k \in K} d_k u_{ik} v_{jk}$$

D is a $K \times K$ diagonal matrix

Can account for both stochastic equivalence and homophily

Cross-sectional network measuring whether an actor indicated that they collaborated with another during the policy design of the Swiss CO_2 act (Ingold 2008)

	Expected Effect	Logit	MRQAP	LSM	ERGM	AME
Conflicting policy preferences						
Business vs. NGO	−	-0.86	-0.87*	-1.37*	-1.11*	-1.37*
Opposition/alliance	+	1.21*	1.14*	0.00	1.22*	1.08*
Preference dissimilarity	−	-0.07	-0.60	-1.76*	-0.44	-0.79*
Transaction costs						
Joint forum participation	+	0.88*	0.75*	1.51*	0.90*	0.92*
Influence						
Influence attribution	+	1.20*	1.29*	0.08	1.00*	1.09*
Alter's influence indegree	+	0.10*	0.11*	0.01	0.21*	0.11*
Influence absolute diff.	−	-0.03*	-0.06*	0.04	-0.05*	-0.07*
Alter = Government actor	+	0.63*	0.68	-0.46	1.04*	0.55
Functional requirements						
Ego = Environmental NGO	+	0.88*	0.99	-0.60	0.79*	0.67
Same actor type	+	0.74*	1.12*	1.17*	0.99*	1.04*

Latent Factor Visualization
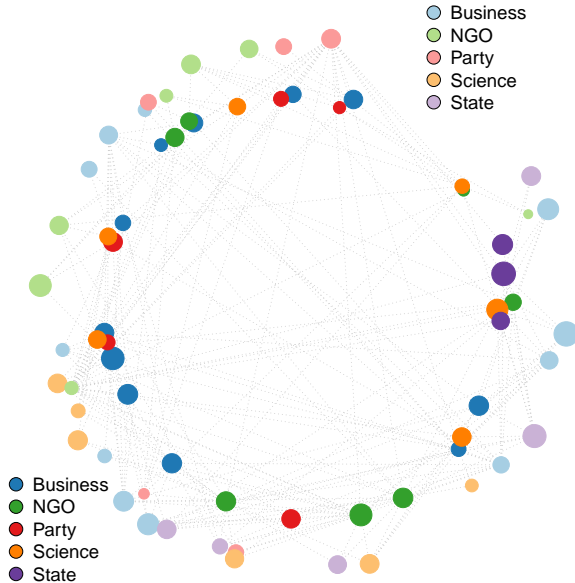
- Randomly divide the $n \times (n-1)$ data points into S sets of roughly equal size, letting s_{ij} be the set to which pair $\{ij\}$ is assigned.
- For each $s \in \{1, \ldots, S\}$:
  - Obtain estimates of the model parameters conditional on $\{y_{ij} : s_{ij} \neq s\}$, the data on pairs not in set s.
  - For pairs $\{kl\}$ in set s, let $\hat{y}_{kl} = E[y_{kl}|\{y_{ij} : s_{ij} \neq s\}]$, the predicted value of y_{kl} obtained using data not in set s.

This procedure generates a sociomatrix of out-of-sample predictions of the observed data
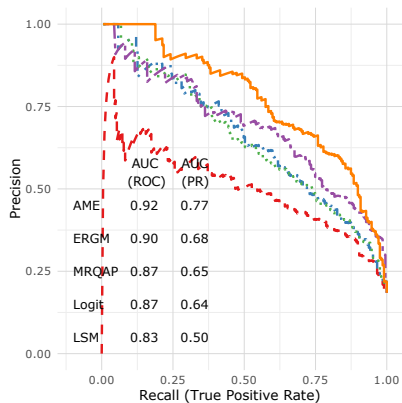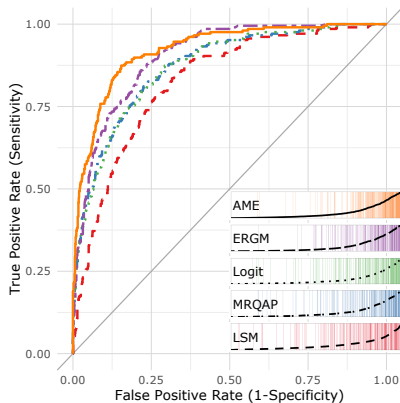
Network Dependencies
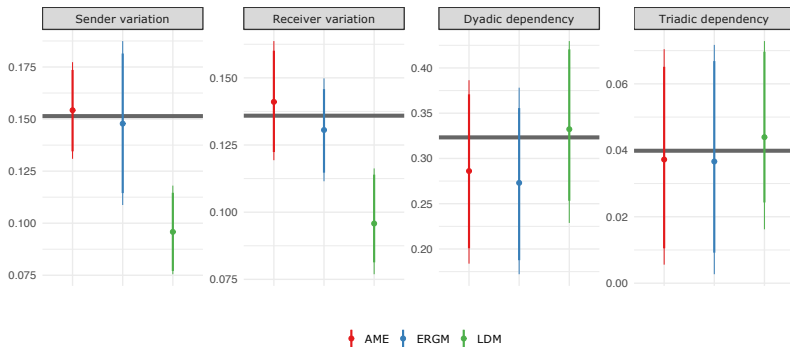
THANKS.

There is a problem with ERGM.

It is a bug, not a feature. If you don't believe us, believe these scholars:

1. Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete expo- nential families. **Journal of the American Statistical Association**, 106(496):1361–1370.

2. Schweinberger, M. and Handcock, M. S. (2015). Local dependence in random graph models: Characterization, properties and statistical inference. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, 77(3):647–676.

3. Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding ex- ponential random graph models. **The Annals of Statistics**, 41(5):2428–2461.

4. Rastelli, R., Friel, N., and Raftery, A. E. (2016). Properties of latent variable network models. **Network Science**, 4(4):1–26.

What Is Really Wrong?

You may assume this is your likelihood surface. Courtesy of Chatterjee and Diaconis, who concluded that sufficient statistics are not sufficient to solve this problem. Estimation of ERGMs contains this flaw, with pseudolikelihood, maximum likelihood, or MCMC.
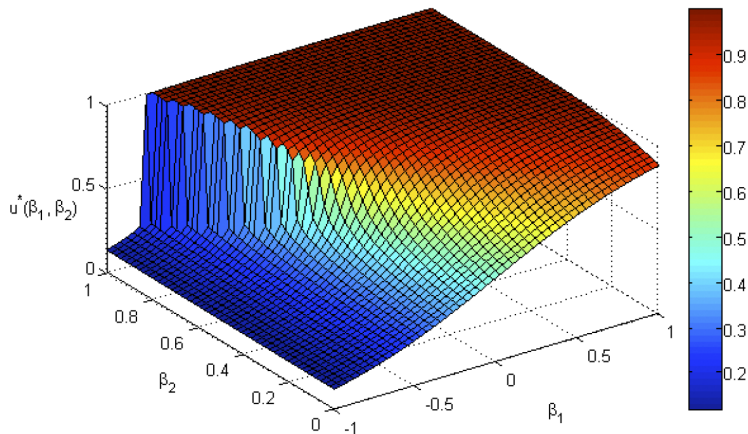
What Does This Mean?

1. Probabilistic ERGM models place almost all of the probability on networks that are either nearly empty (degenerate) with no linkages or nearly saturated with all nodes being interconnected.

2. The likelihood surface contains steep or discontinuous gradients that render it impossible to solve numerically (or analytically). Even (especially) for very small networks this is problematic.

3. Find a different way to estimate network data. (See above!).

Extant Political Science Examples Using Latent Distance Models Are Wrong.

1. There is a problem with the MCMC estimator in `latentnet`. Don't use it until it is fixed.
2. Latent distance models are hard to correctly interpret, in prominent evaluations in the literature it is done incorrectly. This is explained in their introduction a decade ago.
3. The advice in Cranmer et al is wrong: Latent network models are not biased and when correctly implemented out perform ERGM models.
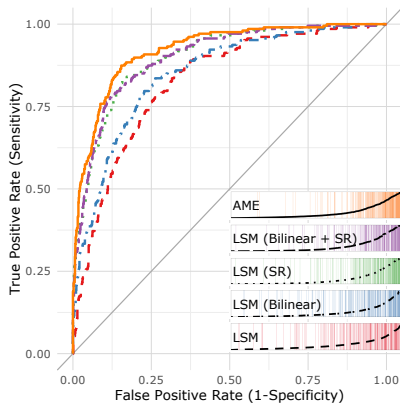
Much current advice is wrong.

Unfortunately some misleading and incorrect advice is in the current literature. Maybe it will be corrected.

1. Cranmer, Leifeld, McClurg, Rolfe (2017 AJPS) stated that latent space models are only preferable if there are few isolates, enough observations, and the interdependence is theoretically uninteresting.

2. All of this is wrong. Not only do they incorrectly assume all latent models to be based on Euclidean distance but their statistical comparisons are based on a program that contains known errors. Moreover, using BIC to choose among different network models is itself questionable.

3. It is not true that "ERGM offers the most preferable combination of model fit and number of parameters," despite their assertions.
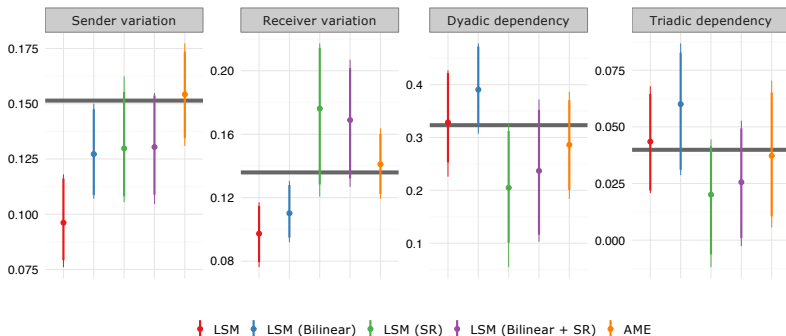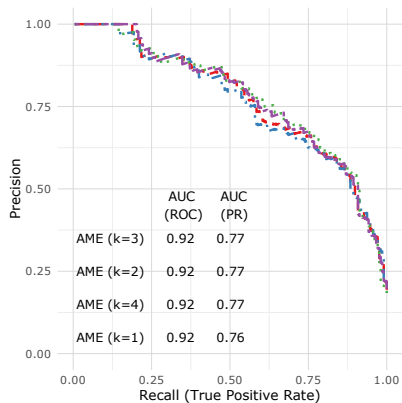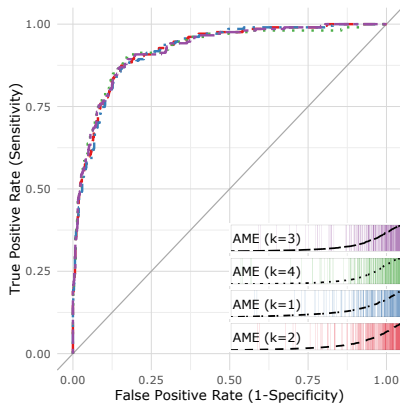
Simulation Comparison
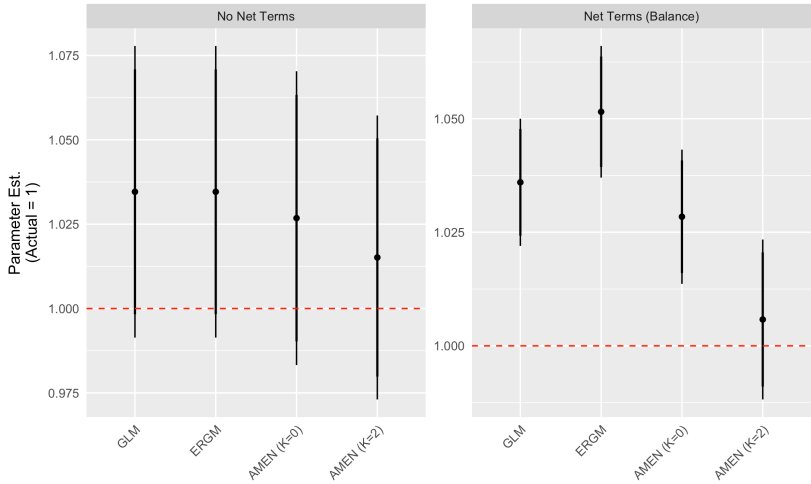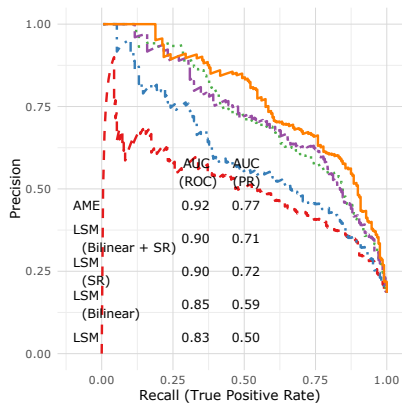
AMEN varying K Net Dependence