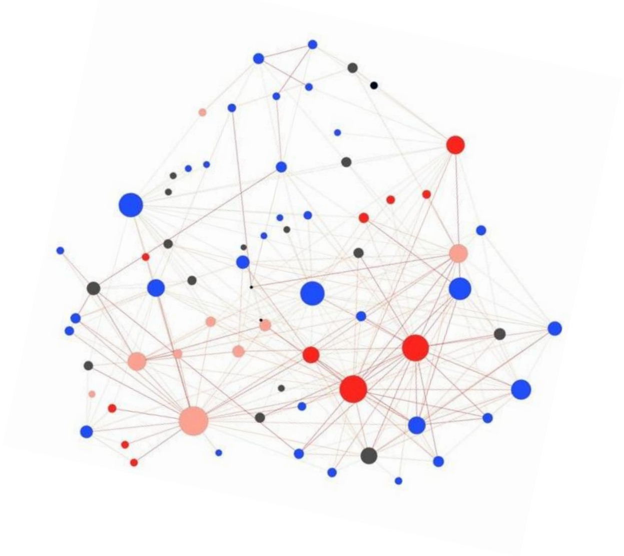


Introduction to inferential network analysis and exponential random graph modeling

Benjamin Rosche – benrosche.com

Social Dynamics Lab workshop @ April 26, 2023

Course materials can be found at:
http://benrosche.com/teaching/isna-workshop/

Outline

Two flavors of inferential social network analysis:

✓ **Network formation: Explain the topology of a network**

✓ Network effects: Explain the beliefs and behaviors of actors embedded in a network

1. Network data
2. Why the dyadic regression model fails
3. Exponential Random Graph Modeling (ERGM)
4. R workshop

Network data

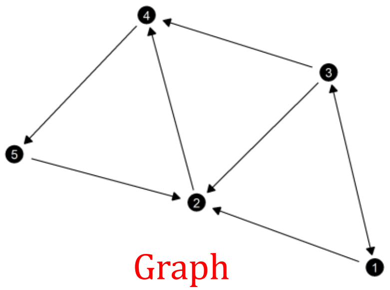
- Any system with potential dependencies between observations can be cast as a network
- Network data representation:

$$w = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Adjacency matrix

from	to	tie
1	2	1
1	3	1
2	4	1
3	1	1
3	2	1
3	4	1
4	5	1
5	2	1

Edge list



Graph

- Sociometric vs egocentric network data
- <u>Two reasons to model dependencies between observations resulting from network embeddedness:</u>
  - Ignoring dependencies leads to bias (i.i.d. assumption)
  - Modeling dependencies explicitly can deepen our theoretical understanding
- <u>Types of explanations:</u>
  - Node-level, edge-level, network-level explanations
  - Exogenous vs endogenous (network-dependent) explanations (e.g., the tendency of people with common friends to become friends)

Why the dyadic regression model fails

- Assumption in regression models: observations are independent and identically distributed (i.i.d.):

$$P(Y) = P(y_1, y_2, \dots, y_n) = P(y_1) \cdot P(y_2) \cdot \dots \cdot P(y_n)$$

- Network dependence: If a change in any y_i causes a change in any y_j, multiplying $P(y_i)$ and P(y_j) does not give the joint probability of the two.
- Network dependence biases both coefficient and uncertainty estimates

Consequences for the dyadic regression model (i.e., logistic regression on dyadic data)

- Bias of the coefficient estimate: it is just a confounding issue
- Bias of the uncertainty estimate: multiplication of observations (with n=100 actors, we have $2\binom{n}{2} = 9900$ possible (directed) dyads)

Rosche - ERGM workshop - April 2023 - Dyadic regression example.R

Exponential random graph models (ERGMs)

The idea in a nutshell:

- Dyadic regression and other dyadic models of networks (e.g., stochastic block models) often do not fit real-world networks very well because of higher-order mechanisms
  - E.g., two nodes who are both linked to the same other node have a higher probability of being friends (triadic closure) than a dyadic model would predict
- ERGMs model the entire networks as an observation of a complex multivariate distribution

ERGM

Network-level specification

θ are the coefficients to be estimated

h(N) are network statistics (akin to explanatory variables in a regression model)

$$P(N, \boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{\theta}^T \boldsymbol{h}(N)\}}{\sum_{N^* \in \mathcal{N}} \exp\{\boldsymbol{\theta}^T \boldsymbol{h}(N^*)\}}$$

Probability that we observe this particular network N

Normalizing constant: the sum of the numerator's value over the set of all possible networks $\mathcal{N}$ (typically all networks with the same node set as the observed network).

ERGM

Dyad-level specification

1/(1+exp(-x))

$$P\left(N_{ij} = 1 \middle| N_{-ij}, \boldsymbol{\theta}\right) = \text{logit}^{-1}\{\boldsymbol{\theta}^T \boldsymbol{\delta_{ij}}(N)\}$$

Probability of an edge from i to j conditional on the rest of the network

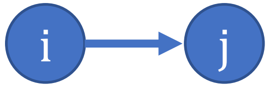
$\boldsymbol{\delta_{ij}}(N)$ is a vector of "change statistics" for each model term. Change statistics record how the network statistics $\boldsymbol{h}(N)$ change if the edge is toggled from off to on while fixing the rest of the network:

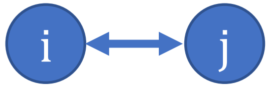
$$\boldsymbol{\delta_{ij}}(N) = \text{h}\left(N_{ij}^{+}\right) - \text{h}\left(N_{ij}^{-}\right)$$

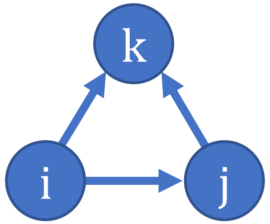
Interpretation:

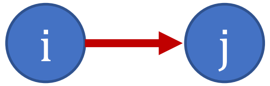
$\delta_{ij}\ \theta$: The contribution of a network statistic to the log-odds of that tie, conditional on all other dyads remaining the same

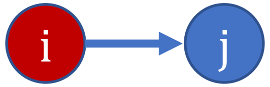
Network statistics h(N)

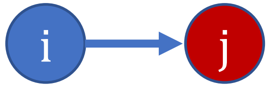
- $h_{\text{edges}} = \sum_{i \neq j} N_{ij}$ 

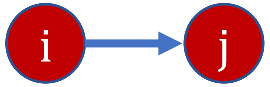
- $h_{\text{mutual}} = \sum_{i \neq j} N_{ij} N_{ji}$ 

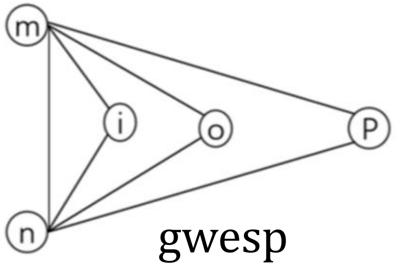
- $h_{\text{triad}} = \sum_{i \neq j \neq k} N_{ij} N_{jk} N_{ki}$ 

- $h_{\text{edgecov}} = \sum_{i \neq j} N_{ij} X_{ij}$ 

- $h_{\text{nodeocov}} = \sum_{i \neq j} N_{ij} X_i$ 

- $h_{\text{nodeicov}} = \sum_{i \neq j} N_{ij} X_j$ 

- $h_{\text{nodematch}} = \sum_{i \neq j} N_{ij} \mathbf{1}(X_i = X_j)$ 



Estimation

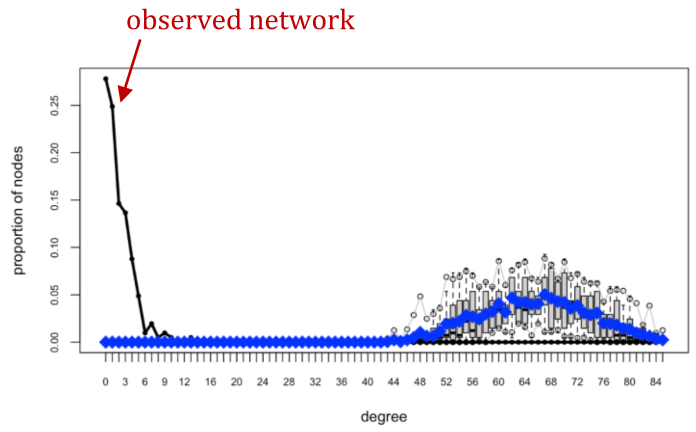
Since the likelihood involves a normalizing constant that is not straightforward to calculate, a combination of Markov Chain Monte Carlo (MCMC) methods to simulate networks and maximum likelihood estimation (MLE) to estimate the parameters is used as soon as endogenous network mechanisms enter the equation.

Idea in a nutshell:

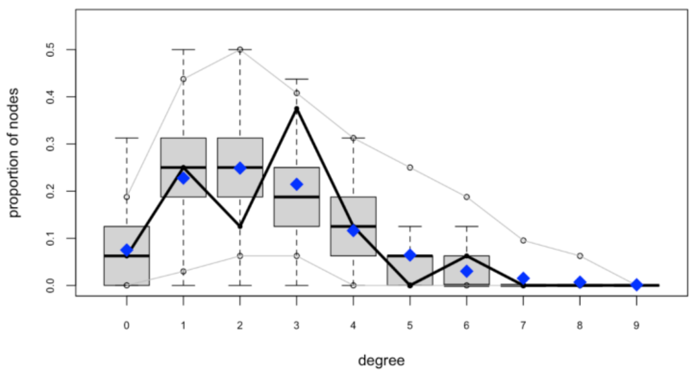
1. Sample networks with (initially random) $\boldsymbol{\theta}$, compute average network statistics (e.g., number of triangles), and check how close these are to the network statistics of the observed network
2. Apply MLE algorithm to (1.) to find a vector $\boldsymbol{\theta}$ that minimizes the distanced between the sample statistics in the sampled and observed network
3. Repeat until parameter estimates provide good model-fit

Checking model-fit and MCMC convergence

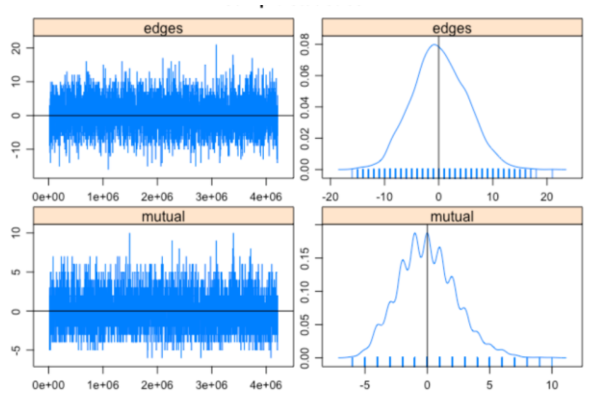
- Many network statistics are inherently correlated, which can lead to poor model-fit (model degeneracy). Therefore, goodness-of-fit needs to be checked.
- ERGMs (with dyad-dependent terms) uses a MCMC algorithm to estimate the parameters. Therefore, the convergence of the algorithm needs to be checked.



Model does not fit the degree distribution well



Model fits the degree distribution better

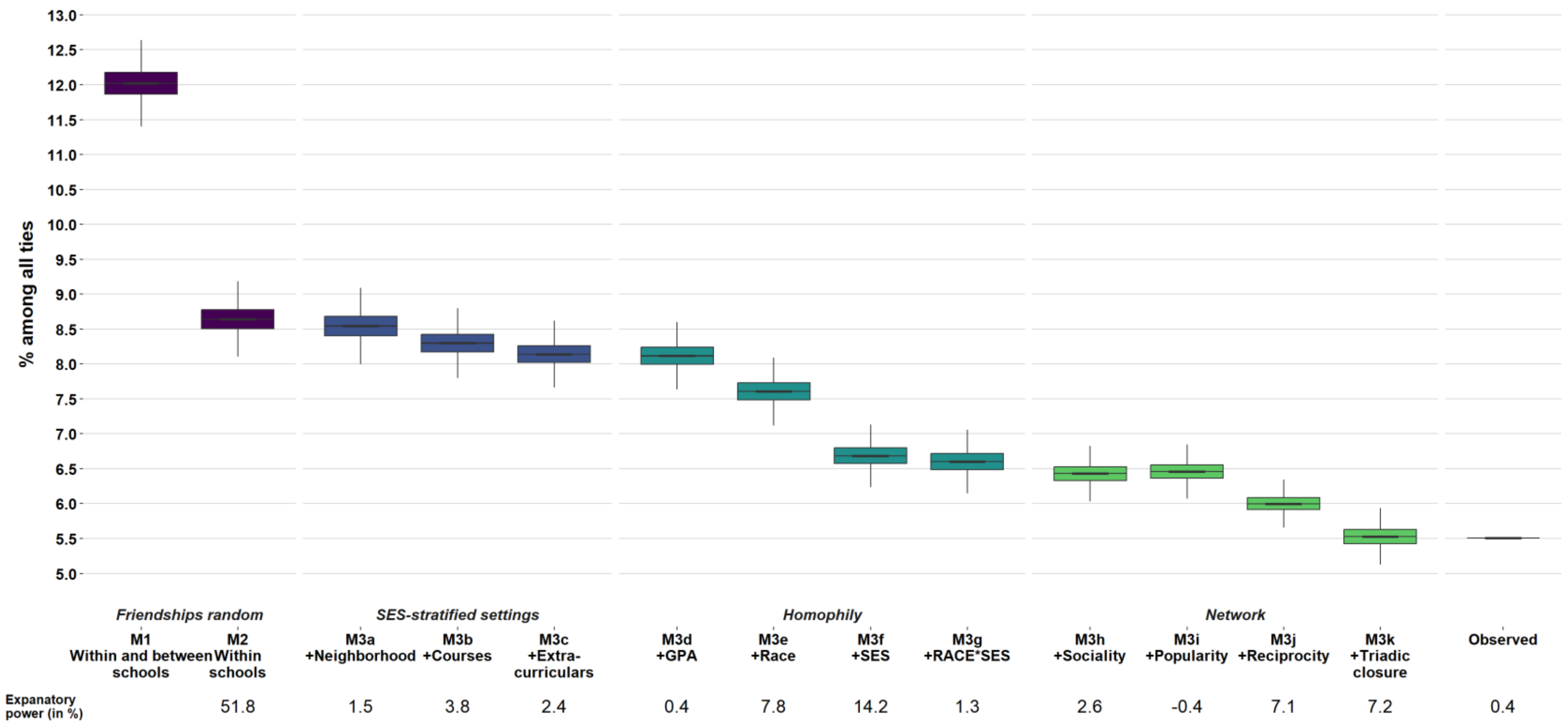


The fat caterpillars in the trace plot suggest the convergence of the algorithm.

Rosche - ERGM workshop - April 2023 - R tutorial.R

Example on how to use ERGMs to examine network structure

Here, I model the determinants of ties that cross socioeconomic boundaries {nodemix("SES")}:



Rosche (forthcoming) - Socioeconomic segregation in friendship networks: Prevalence and determinants of same- and cross-SES friendships in US high schools.

Acknowledgements

This workshop draws on examples and equations from

- Inferential social network analysis course by Philip Leifeld
- statnet.org

Further reading

- Robins, G., Pattison, P., & Woolcock, J. (2005). Small and other worlds: Global network structures from local processes. American Journal of Sociology, 110(4), 894-936.
- Snijders, T. A., & Steglich, C. E. (2015). Representing micro–macro linkages by actor-based dynamic network models. Sociological methods & research, 44(2), 222-271.
- An, W., Beauvile, R., & Rosche, B. (2022). Causal Network Analysis. Annual Review of Sociology, 48, 23-41.
- Duxbury, S. (forthcoming?): A General Framework for Micro-Macro Analysis in Social Networks Research.