Exponential Random Graph (p^*) Models

Janet Xu

Soc Stats Reading Group

October 26th, 2017

Why model social networks?

1. To capture both the regularities and randomness in the social processes that give rise to a network of social ties.
2. To infer whether certain network substructures that might result from social mechanisms (e.g. triadic closure) are more commonly observed in the network than by chance.
3. To evaluate the contribution of different social mechanisms that could produce similar effects (e.g. birds of a feather or friend of a friend?)
4. To understand complexity, like network evolution or multiple network structures
5. To investigate how localized social processes and structures can combine to form global network patterns.

Examples

- Political centralization in Renaissance Florence – how did elites consolidate? (Padgett and Ansell 1993)
- Racial homogeneity among student friendship networks – racial homophily, tendency to befriend friends of friends, homophily on other dimensions, foci? (Goodreau et al. 2009, Wimmer and Lewis 2010)
- Co-sponsorship in Congress – shared committees, shared party, shared state? (Fowler 2006, Cranmer and Desmarais 2010).

The logic behind ERGMs for social networks

- Conceptualize the observed network data as just *one* realization of a set of possible networks with similar important characteristics produced by some unknown stochastic process.
- Similar important characteristics: at the very least, same number of n actors/nodes.
- A statistical model for a network on a given set of actors assigns a probability to all possible networks on those actors.
- The range of possible networks and their probability of occurrence under the model is represented by a probability distribution on the set of all possible graphs.
- Estimate model parameters using observed network as guide.

Five Implicit Steps

1. Assume each tie between two nodes is a random variable: e.g. $y_{ij} = 1$ if there is a tie between i and j and 0 if not.
2. Define the contingencies among the network variables: e.g. transitive triads among close friends, reciprocated ties.
3. These dependent contingencies can translate to a particular model.
4. Constrain the number of parameters, usually through homogeneity assumptions.
5. Estimate and interpret model parameters

General Form

$$Pr(\mathbf{Y} = \mathbf{y}) = (\frac{1}{\kappa}) exp\left\{\sum_{A} \eta_A g_A(\mathbf{y})\right\}$$

- $\mathbf{Y}$ = a possible network that could be observed (matrix of all random variables)
- $\mathbf{y}$ = observed network
- η_A = parameter corresponding to the configuration A
- $g_A(\mathbf{y}) = \prod_{y_{ij} \in A} y_{ij}$ = network statistic corresponding to configuration A
- κ = normalizing quantity

What is a "configuration"?

- Configurations represent possibilities.
- A configuration A refers to a subset of tie variables and corresponds to a small network substructure.
- If a set of possible edges represents a configuration in the model, then the general form equation implies that any subset of possible edges is also a configuration. Thus, single edges are always configurations.
- Could also include reciprocated ties, transitive triads, etc.
- $g_A(\mathbf{y})$ tells us whether the configuration A is in fact observed in the network $\mathbf{y}$.

Dependence assumptions

- We can think of graphs as being generated by potentially overlapping configurations. **But only some configurations are relevant to the model**.
- Dependence assumptions define what these configurations are. The only configurations relevant to the model are those in which all possible ties are **mutually contingent** on each other.
- η_A is zero whenever variables in configuration A are conditionally independent of each other.
- For example, if we assume that the existence of a tie from i to j depends on if j has a tie to i (i.e. reciprocity), then one configuration in the model would be the set of variables $\{Y_{ij}, Y_{ji}\}$ and we would estimate a parameter for this configuration.

Homogeneity Assumption

- Note that the general form equation refers to estimating a parameter for each configuration A. That is so many parameters!!! ($n(n-1)/2$ for reciprocity alone)
- So, impose a homogeneity assumption by equating parameters when they refer to the same type of configuration.
- This assumes that certain regularities are the same for the entire network regardless of which nodes are involved.
- Can also use less strict assumptions to constrain parameters – for example, by equating parameters for isomorphic configurations involving similar types of actors (e.g. girl-girl reciprocity parameter is different from boy-boy reciprocity parameter).

Bernoulli random graph distributions

$$\Pr(\mathbf{Y} = \mathbf{y}) = (\frac{1}{\kappa}) exp(\sum_{i,j} \eta_{ij} y_{ij})$$

- Assume that edges are **independent**
- Under this dependence assumption, the only model-relevant configuration is the single edge
- $g_A(y)$ tells us whether configuration A is observed or not and here every set A is a single possible edge Y_{ij}, so the network statistic is simply y_{ij}

Bernoulli cont.

Strong homogeneity assumption:

$$Pr(\mathbf{Y} = \mathbf{y}) = (\frac{1}{\kappa})exp(\theta L(\mathbf{y}))$$

Here $L(\mathbf{y})$ is the number of arcs (i.e. directed ties) in the graph. θ is the edge or density parameter

Alternative: a priori block structure and block homogeneity

$$Pr(\mathbf{Y} = \mathbf{y}) = (\frac{1}{\kappa})exp(\theta_{11}L_{11}(\mathbf{y}) + \theta_{12}L_{12}(\mathbf{y}) + \theta_{21}L_{21}(\mathbf{y}) + \theta_{22}L_{22}(\mathbf{y}))$$

where $L_{11}(\mathbf{y})$ is the number of arcs within the first block (e.g. if our blocks were boys and girls this could be girl-girl arcs) and $L_{12}(\mathbf{y})$ is the number of arcs from block 1 to block 2 (e.g. girl-boy arcs)

Dyadic Models

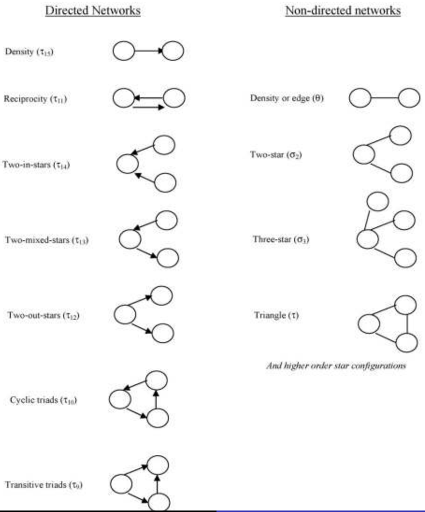
$$Pr(\mathbf{Y} = \mathbf{y}) = (\frac{1}{\kappa})exp(\theta \sum_{i,j} y_{ij} + \rho \sum_{i,j} y_{ij} y_{ji})$$
$$= (\frac{1}{\kappa})exp(\theta L(\mathbf{y}) + \rho M(\mathbf{y}))$$

- Estimate one parameter for edge/density and another for reciprocity (assuming homogeneity)
- Can also condition on node-level attributes by incorporating sender and receiver effects treated as random effects
- My understanding is that stochastic block models are also dyadic?
- Not very realistic because real-world social networks tend to have triangles and other higher-ordered configurations

Markov random graphs

- **Markov dependence** (Frank and Strauss 1986), in which a possible tie from i to j is assumed to be contingent on any other possible tie involving i or j.
- Or, the assumption that two possible network ties are conditionally dependent when they have a common actor.

Markov random graphs cont.



Markov random graphs cont.

Example of Markov RGM for non-directed network with edge, two-star, three-star, and triangle effect parameters (assuming homogeneity for isomorphic configurations):

$$Pr(\mathbf{Y} = \mathbf{y}) = (\frac{1}{\kappa})exp(\theta L(\mathbf{y}) + \sigma_2 S_2(\mathbf{y}) + \sigma_3 S_3(\mathbf{y}) + \tau T(\mathbf{y}))$$

Note that some statistics in the model are higher order to others. For instance, if there are many two-stars present, some triangles will form by chance. But if the triangle effect is larger than by chance, this should be reflected in the model estimate.

Additional Dependence Assumptions

- **Node-level effects**: say, distribution of ties given distribution of attributes: $\Pr(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x})$.
- **Markov attribute assumption**: attribute of i influences possible ties that involve i
- **Setting structures**: Dependencies within social settings, drawing on Feld (1981)'s foci theory. See Pattison and Robins (2002).
- **Realization-dependent models**: non-Markov dependencies among ties that do not share an actor but might be interdependent through third party links
- **New types of homogeneity constraints** such as including higher-order star and triangle effects but constrained with a weighted sum with alternating signs

Old Approach to Estimation: Pseudo-Likelihood

- Generally bad.
- Pro: relatively easy to fit even complicated models
- Con: parameter estimates may be biased
- Con: standard errors are approximate at best; may be too small
- Con: properties not well understood, e.g. cannot assume that pseudo-likelihood deviance is asymptotically distributed like Chi-squared, should not rely on Wald statistic as a means to decide whether a parameter is significant or not.
- Con: misleading estimates for near degenerate models

Pseudo-Likelihood

Done by transforming the general form equation into a conditional form:

$$log\left[\frac{Pr(Y_{ij}=1|\mathbf{y}_{ij}^{C})}{Pr(Y_{ij}=0|\mathbf{y}_{ij}^{C})}\right]=\sum_{A(Y_{ij})}\eta_A d_A(\mathbf{y})$$

- $d_A(\mathbf{y})$ is the change in the value of the network statistic $z_A(\mathbf{y})$ when y_{ij} goes from 1 to 0.
- $\mathbf{y}_{ij}^{C}$ is all the observations of ties in observed graph except for y_{ij} – so holding all other ties constant
- This looks like a logistic regression but **is not** a logistic regression because we explicitly do not make independence assumptions!

Better Approach: Markov Chain Monte Carlo MLE

1. Simulate a distribution of random graphs from starting set of parameter values
2. Subsequently refine of parameter values by comparing distribution of graphs against observed graph
3. Repeat until parameter estimates stabilize

- But, still problems: **near degeneracy** occurs when a model implies that only a few graphs had anything other than very low probability
- The estimation process does not converge and we can not obtain consistent parameter estimates with MCMCMLE. Markov graph models might be inappropriate for the data.

Cool New Developments

- Actor-oriented models
- More interest in non-Markovian assumptions
- Hidden Markov Models
- Better estimation strategies?