36-780: Social Network Modeling

ERGMs as models

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Outline

- Special cases of ERGMs
  - Erdos-Renyi-Gilbert model
  - Beta model
  - p_1 model
- Exponential Random Graph Models (ERGMs)
  - Computational explosion
  - Independent-dyads approximation
  - MCMC-MLE
- ERGMs: The Good, The Bad, The R Example
- HW02, for Tuesday, is on line

The Erdos-Renyi-Gilbert Model

- G(N,p) specifies a model for a directed graph G on N nodes, with iid probability p of an edge y_{ij} between any two nodes i and j.

  Then $\#E(G) = K \sim Binomial\left(\binom{N}{2}, p\right)$

- G(N,K) is the equivalent "hypergeometric-like" model, that conditions on the number of edges K, with p = $K/\binom{N}{2}$.

The *beta* model for social networks

- The E-R-G model assumes Y_{ij} iid with

$$\log \frac{P(Y_{ij})}{1-P(Y_{ij})} \equiv \theta$$

- The *beta model* assumes Y_{ij} indepenent with

$$\log \frac{P(Y_{ij})}{1-P(Y_{ij})} = \beta_i + \beta_j$$

  This allows for very simple variation in degree across nodes

The p_1 model (Holland & Leinhardt 1981)

- Easiest to state in terms of the joint likelihood for the whole adjacency matrix Y:

$$\log P(Y = y) \propto \theta y_{++} + \sum_i \alpha_i y_{i+} + \sum_j \beta_j y_{+j} + \rho \sum_{ij} y_{ij} y_{ji}$$

  - *A subscript + means "sum over that index"*
  - The sufficient statistics are the out-degrees y_{i+} and in-degrees y_{+j} for each node, and the number of "reciprocal dyads" $\sum_{ij} y_{ij} y_{ji}$
  - θ is an overall tie propensity
  - α_i is node i's "gregariousness" or "expansiveness"
  - β_j is node j's "attractiveness" or "popularity"
  - ρ is the tendency to reciprocate ties in the graph.

Exponential Random Graph Models (p^* models[1]); Frank & Strauss (1986)

- In the p_1 model, the term

$$\rho \sum_{ij} y_{ij} y_{ji}$$

  counts a certain "topological feature" (reciprocal ties).
- Could replace $\sum_{ij} y_{ij} y_{ji}$ with a summary *g(y)* of some other topological feature(s), getting a new model term

$$\theta^* g(y)$$

- This may depend on more than one edge at a time, and hence the model no longer has independent dyads. E.g., $g(y) = \sum_{i,j,k} y_{ij} y_{jk} y_{ki} = \#$ of 3-cycles.

[1]Anderson, Wasserman & Faust (1992) use this name. See also Anderson, Wasserman & Crouch, (1999) .

Exponential Random Graph Models (ERGMs)[2]

- This leads to a model of the form

$$\log P(Y=y) \propto \theta y_{++} + \sum_i \alpha_i y_{i+} + \sum_j \beta_j y_{+j} + \theta^* g(y)$$

- No need to retain the p_1 terms, and we could have more than one *g(y)*; this leads to

$$P(Y=y) = \frac{\exp\{\vec{\theta}^{\mathsf{T}} \vec{g}(y)\}}{\kappa(\vec{\theta}, \mathcal{Y})}$$

where $\vec{\theta}$ is a vector of parameters, $\vec{g}(y)$ is a vector of summary statistics, $\kappa(\vec{\theta}, \mathcal{Y})$ is a normalizing constant, summed over the set $\mathcal{Y}$ of "all" networks on *E(Y)*.

[2]Following, more or less, Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, *24*(3). [http://www.jstatsoft.org/v24/i03/]

ERGM computational explosion…

- The normalizing constant

$$\kappa(\vec{\theta}, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\vec{\theta}^{\mathsf{T}} \vec{g}(y)\}$$

is a sum over all possible graphs in $\mathcal{Y}$ and this can be a big number:

$$\text{card}(\mathcal{Y}) = 2^{\binom{N}{2}}$$

- Usually infeasible to compute the MLE directly. Instead, we see
  - Independent-dyad approximation (Strauss & Ikeda, 1990)
  - MCMC-MLE (Snijders, 2002)

ERGM: Independent-dyads approx.

- Let Y_{ij} indicate (1 or 0) an edge, let Y^c_{ij} be the rest of the graph, and let y^+ and y^- be the graph with y_{ij}=1, and y_{ij}=0, respectively. Then

$$\log \frac{P[Y_{ij}=1|Y^c_{ij}=y^c_{ij}]}{P[Y_{ij}=0|Y^c_{ij}=y^c_{ij}]} = \vec{\theta}^{\mathsf{T}}[\vec{g}(y^+) - \vec{g}(y^-)]$$

- If we now "pretend" that $\vec{g}(y)$ does not depend on Y^c_{ij}, we get an independent dyads model

$$\log \frac{P[Y_{ij}=1]}{P[Y_{ij}=0]} = \vec{\theta}^{\mathsf{T}}[\vec{g}(y_{ij}|_{y_{ij}=1}) - \vec{g}(y_{ij}|_{y_{ij}=0})]$$

that is easy to estimate (e.g., glm).

ERGM: Full (MCMC) approx MLE (Snijders, 2002)

- The "indep dyads" approximation made ERGMs feasible and hence popular
  - But it can produce really bad answers (van Duijn, Gile, and Handcock 2007)
- Snijders (2002) exploits a trick of Geyer & Thompson (1992) to approximate (a ratio involving) $\kappa(\vec{\theta}, \mathcal{Y})$ with MCMC, and hence to get approximate full MLEs.
- *All of the above also works when* $\vec{g}(y, X)$ *depends on covariates X!*

Building ERGMs…

- What can be $\vec{g}(y, X)$ terms in

$$\begin{aligned} \log P(Y=y) &\propto \vec{\theta}^{\mathsf{T}} \vec{g}(y, X) \\ &= \theta_1 g_1(y, X) + \theta_2 g_2(y, X) + \cdots + \theta_k g_k(y, X) \quad ? \end{aligned}$$

  - $g(y, X) = y_{i+},\ y_{+j},\ y_{++},\ \sum_{ij} y_{ij} y_{ji},\ \sum_{ijk} y_{ij} y_{jk} y_{ki}$, etc.
  - $g(y, X) =$ the 16 counts in a triad census
  - $g(y, X) = 1$ if node i is male, 0 else
  - $g(y, X) = 1$ if nodes i and j have same gender, 0 if not.
  - $g(y, X) =$ difference in income between i and j
  - Etc. etc. etc.!

ERGMs in R: The Good, The Bad, …

- The above is the GOOD news about ERGMs.
- We will look at an example in R.
- You will also look at the “ergm” package in your hw.

- The BAD news about ERGMs will come next week (and, somewhat, in the R example…)

Outline

- Special cases of ERGMs
  - Erdos-Renyi-Gilbert model
  - Beta model
  - p_1 model
- Exponential Random Graph Models (ERGMs)
  - Computational explosion
  - Independent-dyads approximation
  - MCMC-MLE
- ERGMs: The Good, The Bad, The R Example
- HW02, for Tuesday, is on line