Exponential Random Graph Models for Social Networks

ERGM Introduction

Martina Morris
Departments of Sociology, Statistics
University of Washington

Carter T. Butts
Departments of Sociology, Statistics, and EECS, and Institute for Mathematical Behavioral Sciences
University of California, Irvine

SUNBELT 15, Brighton
From Description to Modeling

- Ultimately, want to do more than describe networks
- **Network modeling**: predict the formation and structure of social networks
- Many examples
  - Conditional uniform graphs, Bernoulli graphs
  - Holland and Leinhardt's $p_1$
  - Degree distribution models, growth models, etc.
- **ERGM**: a general representation for such models
  - Draws on theory of statistical exponential families
  - Not really a "type" of model (in a scientific sense), but a way of representing and working with new and existing models!
Initial Intuition: Factors in Tie Formation

- All ties are not equally probable
  - Chance of an (i,j) edge may depend on properties of i and j
  - Can also depend on other (i,j) relationships

- Some examples:
  - Homophily
  - Propinquity
  - Multiplexity
AddHealth Friendship Network, by Grade
Logistic Network Regression

• A classic starting point: why not treat edges as independent, w/log-odds as a linear function of covariates?
  - Special case of standard logistic regression
  - Dependent variable is a network adjacency matrix

• Model form:

\[
\log\left( \frac{\Pr(Y_{ij} = 1)}{\Pr(Y_{ij} = 0)} \right) = \theta_1 X_{ij1} + \theta_2 X_{ij2} + \ldots + \theta_m X_{ijm} = \theta^T X_{ij}
\]

  - where \( Y_{ij} \) is the value of the edge from \( i \) to \( j \) on the dependent relation, \( X_{ijk} \) is the value of the \( k \)th predictor for the \((i,j)\) ordered pair, and \( \theta_1, \ldots, \theta_m \) are coefficients

  • \( \log(p/(1-p)) = \text{logit}(p) \), maps \((0,1)\) to \((-\infty, \infty)\)
Moving Beyond the Logistic Case

• The logistic model can be quite powerful, but still very limiting
  – No way to model conditional dependence among edges
    • E.g., true triad closure bias, reciprocity
  – Cannot handle exotic support constraints
    • What if your network must be transitive (e.g., sports contests, entailments), an interval graph (e.g., life history graphs), etc?

• A more general framework: discrete exponential families
  – Very general way of representing discrete distributions
  – Turns up frequently in statistics, physics, etc.
Beyond Independence: the Star Terms

- **Simple subgraph census terms**
  - $k$-stars: number of subgraphs isomorphic to $K_{1,k}$
  - $k$-in/out/mixed-stars: number of subgraphs isomorphic to orientations of $K_{1,k}$

- **Interpretations**
  - Tendency of edges to “stick together” on endpoints (“edge clustering”)
  - Fixes moments of the degree distribution
    - 1-stars fix mean degree, 2-stars fix variance
Another Way to See Stars:
Degree Terms

- Natural reparameterization of the star terms
  - $i$th degree term: number of vertices of degree $i$
    - Likewise for indegree, outdegree terms
  - Can be derived from the full set of star terms (and vice versa)

- Interpretation
  - Non-parametric model for the degree distribution
  - Note: do not confuse with sender/receiver terms!
    - Latter refer to specific vertices, do not create dependence among edges

$d_0=0$, $d_1=5$, $d_2=2$, $d_3=0$, $d_4=2$, $d_5=1$, $d_6=0$, $d_7=0$, $d_8=0$, $d_9=0$
Triad Census Terms

- Most basic terms for endogenous clustering
  - Each term counts subgraphs isomorphic to triads of a given type (i.e., elements of the triad census)
  - In practice, triangles, cycles, and transitives most often used

- Interpretation
  - Tendencies towards transitive closure, cycles, etc.
  - Transitivity can be an indicator of latent hierarchy
  - Cyclicity can be an indicator of extended reciprocity
Mc, MC, and MCMC In One Slide

- **Markov chain**
  - Stochastic process $X_1, X_2, \ldots$ on state space $S$, such that $p(X|X_{i-1}, X_{i-2}, \ldots) = p(X_i|X_{i-1})$ (i.e., only the previous state matters – this is the Markov condition)

- **Monte Carlo procedure**
  - Any procedure which uses randomization to perform a computation, having a fixed execution time and uncertain output (compare w/Las Vegas procedures)

- **Markov chain Monte Carlo (MCMC)**
  - Family of procedures using Markov chains to perform computations and/or simulate target distributions; often, these cannot be done any other way

- **Important Example: Metropolis Algorithm**
  - Given $X_i$, draw $X'$ from $q(X_i)$; w/probability $\min(1, p(X')/p(X_i))$, let $X_{i+1} = X'$, else let $X_{i+1} = X_i$. Repeat for $i+1, i+2$, etc.
  - Started w/arbitrary $X_0, X_0, X_1, \ldots X_n$ converges to $p(X)$ in distribution as $n \to \infty$
  - Requires some constraints on $q$, but is very general – used when we can't sample from target distribution $p$ directly (as when $p$ is an ERG distribution)
Example: floboxbusiness w/edges
Example: flobusiness w/edges
flobusiness w/edges, marriage
flobusiness w/edges, marriage
flobusiness w/edges, marriage, isolates (0-degree)
Final flobusiness Model

- We now have a model in which we can be reasonably confident (MCMC diags also OK):

```
Formula: flobusiness ~ edges + edgecov(flomarriage) + degree(0)
Newton-Raphson iterations: 3
MCMC sample of size 10000

Monte Carlo MLE Results:

        Estimate Std. Error    MCMC s.e. p-value
edges  -1.7274     0.2887    0.023 < 1e-04 ***
edgecov.flomarriage  2.4852     0.4554    0.041 < 1e-04 ***
degree0   2.3828     0.7141    0.061   0.0114 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null Deviance: 166.355 on 120 degrees of freedom
Residual Deviance: 64.377 on 117 degrees of freedom
Deviance: 101.978 on  3 degrees of freedom

AIC: 70.377  BIC: 78.74
```