36-780: Social Network Modeling

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

Class Materials

- In Class
  - These notes
- On web (http://www.stat.cmu.edu/~brian/780)
  - Class notes, handouts, links, etc.
  - Homework
  - Reading
  - Computing
- On Blackboard (blackboard.andrew.cmu.edu)
  - Discussion Board
  - Turn in written assignments (pdf!) when needed

Outline

- Introduction & office hours
- Syllabus Stuff
- Social Networks…
  - ❑ Descriptive analysis in R
  - ❑ Erdos-Renyi-Gilbert model
  - ❑ P_1 model
  - ❑ P_2 model
- Directions from here…
- HW01 is posted online (two due dates this week!)

Introduction – about us

- Instructor
  Brian Junker
  132E Baker Hall
  (412) 268-8874
  brian@stat.cmu.edu
- Office Hours
  - ❑ 132E Baker
  - ❑ Tues at Noon
  - ❑ Thurs at 1:30pm
- TA
  Mauricio Sadinle
  Wean 8115
  (412) 268-5610
  msadinle@stat.cmu.edu
- Office Hours
  - ❑ ___________
  - ❑ ___________
  - ❑ ___________

Syllabus Stuff – course materials

- *No textbook*
  - Just class notes, web materials, journal articles
- *If you are interested, these two books summarize traditional material well*
  - de Nooy, W., Mrvar, A., & Batagelj, V. (Eds.). (2005). *Exploratory social network analysis with Pajek (Vol. 27).* Cambridge University Press.
  - Kolaczyk, E. D. (2009). *Statistical analysis of network data*. Springer
- *This long article surveys much of the current state of affairs*
  - Goldenberg, A., Zheng, A. X., Fienberg, S. E., & Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning, 2(2),* 129–233.
- *And some of the newest stuff we'll be discussing is here*
  - http://hnm.stat.cmu.edu

My goals for the course

- Passing understanding of descriptive analysis of social network data.
- Understand how a generative model entails a statistical model, and how statistical models offer avenues for
  - combining analyses across ensembles of networks,
  - extending analyses from smaller samples to larger ones, etc.
- Engage some current research questions in social network analysis.
- Apply what you have learned to a small project.

Main Computing Tools

- R
  - Great “breadboarding” system
  - Can do moderately large problems
  - Big open source community
  - Mostly not set up for big data
- A little “meatball programming” experience also helpful
- If you want to use something else, you may, but please document what you are doing

What you will do in the course

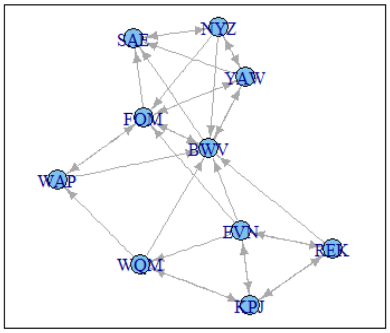
- Throughout the mini:
  - Post questions and answers on Blackboard
  - Participate in class
- First ca. 2/3 of mini:
  - Computer labs
  - Read papers, discuss in class
- Last ca. 1/3 of mini:
  - Everyone presents 1-2 papers, or presents a small project

Paper presentation or project?

- *If you have a projec*t on social network modeling that is a part of your research, I encourage you to make that be your project for the mini.
- *If you do not* have a network data analysis or modeling project, then you should select one or more research papers to present to the class.
- *In either case* (paper(s) or project), you will
  - ❑ Lead a class discussion on your paper(s) or project, using projected slides or other tools as appropriate.
  - ❑ Write a short "conference paper" on your work.

Social Networks

- Nodes, vertices
- Edges, links, ties
- Egos vs alters
- Directed vs Undirected
- Node attributes
- Edge attributes
- Graph, sociogram
- Adjaceny matrix, weight matrix, sociomatrix



	KPJ	WQM	FOM	SAE	NYZ	YAW	EVN	BWV	WAP	REK
REK	■						■	■		
WAP			■					■		
BWV			■	■		■				
EVN	■	■	■					■		■
YAW			■	■	■			■		
NYZ			■	■		■		■		
SAE					■					
FOM				■		■		■	■	
WQM	■							■	■	
KPJ		■					■			■

Descriptive analysis often emphasizes topological features, e.g.:

- **Node Centrality**
  - Degree centrality (in-degree, out-degree)
  - Closeness
    - average geodesic distance to get from/to this node, to/from any connected node
  - Betweenness
    - Fraction of geodesic paths passing through this node
- **Edge Centrality** similar (esp. betweenness)
- **Block or community structure**
- **Other topological features** (triads, stars, cliques…)
- **When there are other covariates**, homophily and similar concepts come into play as well

Digression to R…

Some basic notation & models[1]

- G = a graph or network;
  - V(G) = its vertices (nodes),
  - E(G) = its edges (ties),
  - and for now N(G) = #V(G), K(G)=#E(G).
- For i, j $\in$ V(G), let y_{ij} be the indicator

$$y_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E(G) \\ 0 & \text{else} \end{cases}$$

- The adjacency matrix is y=A(G).
- If the edges have weights, then y_{ij} will have weights as values instead

[1] Most of the remaining material in this lecture was taken from Fienberg, S.E. (2012). A Brief History of Statistical Models for Network Analysis and Open Challenges. *Journal of Computational and Graphical Statistics, 21:4,* 825-839. http://dx.doi.org/10.1080/10618600.2012.738106

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 = $K/\binom{N}{2}$.

Changepoint properties of the E-R-G model, depending on λ=Np

- If λ<1, then a graph generated from G(N,p) will have no connected components larger than O(log N), a.s., as N→∞.
- If λ=1, then a G(N,p) graph will have a largest compnent of size O($N^{2/3}$) , a.s., as N→∞.
- If λ→c>1 as N→∞, then a G(N,p) graph will have a unique "giant" component containing a positive fraction of nodes, and no other component will be larger than O(log N).

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.

The p_2 model (Snijders et al, 2000's)

- The principal addition of p_2 over p_1 is to allow covariates X_1, X_2, … and random effects A, B… in modeling the α's and β's, e.g.

$$\begin{aligned} \vec{\alpha} &= \mathbf{X}_1 \vec{\gamma}_1 + \vec{A} \\ \vec{\beta} &= \mathbf{X}_2 \vec{\gamma}_2 + \vec{B} \end{aligned}$$

  where the arrows denote vectors of parameters or random effects, and X_1 and X_2 are covariate matrices.

  - This can be extended to other parameters, and can be extended to model multiple social networks.

Directions from here…

- A variation on the E-R-G model called the *Exchangeable Random Graph Model* is a simple way to probabilistically model block/community structure
- The p_1 model is also a natural precursor to the general *p^* models*, also known as *Exponential Random Graph Models (ERGMs)*. We will discuss them in the next lecture or two…
- p_1 is also a precursor to *conditionally independent dyad (CID)* models.
- The p_2 model and its generalization by Zijlstra is a precursor to *Hierarchical Network Models (HNMs)*.

Summary

- Introduction & office hours
- Syllabus Stuff
- Social Networks…
  - Descriptive analysis in R
  - Erdos-Renyi-Gilbert model
  - P_1 model
  - P_2 model
- Directions from here…
- HW01 is posted online (two due dates this week!)