Workshop on Social Networks in Education Research

Introduction to Social Networks
Tracy Sweet, Brian Junker, Andrew Thomas
http://hnm.stat.cmu.edu

This work is supported in part by IES (US Dept of ED) Award #R305D120004 and NSF Award #SES-1229271.

Our Goals for the Workshop

- Introduce you to descriptive analysis of social network data.
- Show you how statistical models can help organize and focus your analysis of social network data.
- Get you excited about intervention studies on social networks in your research.
- Get you excited about applying CIDnetworks and Hierarchical Network Models (HLSM) to your data.
- Bring you into our social network, so you can help us make CIDnetworks and Hierarchical Network Models better over time, with your suggestions & ideas.
Examples of Social Networks Online

- Facebook
  - Friending is a “symmetric” relationship between people
  - Posting on someone’s wall is asymmetric
  - Following is asymmetric

- Twitter
  - Following and retweeting are both asymmetric

- Classroom 2.0

Examples of Social Networks in Education Research

- **Bully Prevention**: Which interventions are effective in reducing bullying behavior? How do they affect the friendship networks themselves? (Dorothy Espelage, Univ. of Illinois, multiple school-level networks)

- **Organizational Structure**: How do org structure and individual factors shape professional advice-seeking networks? (Jim Spillane, Northwestern, 30 school networks)

- **Evolving Friendship Ties**: Friendship data was collected on 5th graders several times over a school year. What factors affect friendships and changes in friendship over time? (Rebecca Madill, Penn State, 25 networks).
The Pitts & Spillane (2009) Data

- School staffing survey given to teachers in 15 schools
  - Pre-K through 8, private and public schools
  - Does teacher i seek advice from teacher j?
  - Demographics, beliefs, and professional experience were also collected:

<table>
<thead>
<tr>
<th>For teachers:</th>
<th>For pairs of teachers (dyads):</th>
<th>For schools:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Years teaching</td>
<td>• Similar # of years in school?</td>
<td>• Catholic?</td>
</tr>
<tr>
<td>• Sense of trust</td>
<td>• Same innovative attitudes?</td>
<td>• School size</td>
</tr>
<tr>
<td></td>
<td>• Teach same grade?</td>
<td></td>
</tr>
</tbody>
</table>

For teachers:
- Years teaching
- Sense of trust

For pairs of teachers (dyads):
- Similar # of years in school?
- Same innovative attitudes?
- Teach same grade?

For schools:
- Catholic?
- School size


An Advice Network

- Teachers are *nodes* or *vertices* in the network

- Teacher i seeks advice from teacher j iff there is an *edge* i -> j
  (*edges=links=ties*)
  - A *dyad* is a pair of nodes; may have an edge or not
  - Advice-seeking is *asymmetric, directed*
  - The graph is a *sociogram*

- Egos vs alters
  - An “*ego*” is the teacher you are looking at right now
  - The “*alters*” are his/her neighbors in the graph
An Advice Network

- The social network can also be represented as a sociomatrix (adjacency matrix, weight matrix)

Some basic notation

- $G$ = a graph or network;
  - $V(G)$ = its vertices (nodes),
  - $E(G)$ = its edges (ties),
  - $N(G)$ = $\#V(G)$, $K(G)$ = $\#E(G)$.
- 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

---

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

- **Graph Density** (fraction of total possible edges in G)
- **Node Centrality Measures:**
  - **Node degree** (how many edges go into or out of this node)
  - **Closeness**
    - \( \frac{1}{\text{average geodesic distance to get from/to this node, to/from any connected node}} \)
  - **Betweenness**
    - Average number of geodesic paths passing through this node
- **Edge Centrality** similar (esp. betweenness)
- **Block or community structure**
- **Other topological features** (triads/transitivity, stars, cliques…)
  - (we will mostly omit these)

For our little network...

- Tie density is
  \[ \frac{K}{N(N-1)} = \frac{32}{90} = 0.36 \]
- Node centrality measures:

<table>
<thead>
<tr>
<th></th>
<th>KPJ</th>
<th>WQM</th>
<th>FOM</th>
<th>SAE</th>
<th>NYZ</th>
<th>YAW</th>
<th>EVN</th>
<th>BWV</th>
<th>WAP</th>
<th>REK</th>
</tr>
</thead>
<tbody>
<tr>
<td>In degree</td>
<td>3.00</td>
<td>2.00</td>
<td>5.00</td>
<td>4.00</td>
<td>2.00</td>
<td>3.00</td>
<td>2.00</td>
<td>7.0</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Out degree</td>
<td>3.00</td>
<td>3.00</td>
<td>4.00</td>
<td>1.00</td>
<td>4.00</td>
<td>4.00</td>
<td>5.00</td>
<td>3.0</td>
<td>2.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.08</td>
<td>0.1</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Betweenness</td>
<td>2.75</td>
<td>3.08</td>
<td>8.75</td>
<td>3.50</td>
<td>4.00</td>
<td>3.50</td>
<td>4.33</td>
<td>12.5</td>
<td>0.50</td>
<td>1.08</td>
</tr>
</tbody>
</table>

- Edge centrality:
  - We show it on the next slide
  - Edges or nodes with high “betweenness” might be on paths between blocks or clusters in the network...
Edge betweenness...

- Average number of geodesics passing through this edge

<table>
<thead>
<tr>
<th>Edge</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPJ -&gt; WQM</td>
<td>3.83</td>
</tr>
<tr>
<td>KPJ -&gt; EVN</td>
<td>4.83</td>
</tr>
<tr>
<td>KPJ -&gt; REK</td>
<td>3.08</td>
</tr>
<tr>
<td>WQM -&gt; KPJ</td>
<td>3.00</td>
</tr>
<tr>
<td>WQM -&gt; BWV</td>
<td>5.58</td>
</tr>
<tr>
<td>WQM -&gt; WAP</td>
<td>3.50</td>
</tr>
<tr>
<td>FOM -&gt; SAE</td>
<td>3.38</td>
</tr>
<tr>
<td>FOM -&gt; YAW</td>
<td>3.38</td>
</tr>
<tr>
<td>FOM -&gt; BWV</td>
<td>1.00</td>
</tr>
<tr>
<td>FOM -&gt; WAP</td>
<td>6.00</td>
</tr>
<tr>
<td>SAE -&gt; NYZ</td>
<td>8.50</td>
</tr>
<tr>
<td>NYZ -&gt; FOM</td>
<td>4.00</td>
</tr>
<tr>
<td>NYZ -&gt; SAE</td>
<td>1.00</td>
</tr>
<tr>
<td>NYZ -&gt; YAW</td>
<td>2.00</td>
</tr>
<tr>
<td>YAW -&gt; BWV</td>
<td>2.00</td>
</tr>
<tr>
<td>YAW -&gt; FOM</td>
<td>2.00</td>
</tr>
<tr>
<td>KPJ -&gt; SAE</td>
<td>1.00</td>
</tr>
<tr>
<td>KPJ -&gt; YAW</td>
<td>4.50</td>
</tr>
<tr>
<td>BWV -&gt; SAE</td>
<td>7.12</td>
</tr>
<tr>
<td>BWV -&gt; NYZ</td>
<td>4.00</td>
</tr>
<tr>
<td>BWV -&gt; YAW</td>
<td>7.12</td>
</tr>
<tr>
<td>WAP -&gt; FOM</td>
<td>3.00</td>
</tr>
<tr>
<td>WAP -&gt; BWV</td>
<td>2.50</td>
</tr>
<tr>
<td>REK -&gt; KPJ</td>
<td>1.75</td>
</tr>
<tr>
<td>REK -&gt; EVN</td>
<td>2.50</td>
</tr>
<tr>
<td>REK -&gt; BWV</td>
<td>5.83</td>
</tr>
</tbody>
</table>

Block or Community Structure

- Edges with high edge-betweenness might be connecting communities (E-B communities)
- A random walk of, say, 4 steps, should get stuck in a community (walktrap communities)
Digression to R…

- Pick one or more schools, and explore the features we have been talking about with it/them. Make some comparisons!

Models for Social Networks

- We will (mostly) skip over classic social network models
  - $P_1$ models
  - $P_2$ models
  - $P_*$ or “Exponential Random Graph Models” (ERGMs)

- Instead we concentrate on scalable generative models:
  - *Dyadic independence models* with covariates
  - *Conditionally independent dyad (CID) models*
  - *Hierarchical Network Models (HNMs)*
Dyadic Independence Models with Covariates

- For adjacency matrix $Y = [Y_{ij}]$, 

\[
\logit P[Y_{ij} = 1] = X_{ij} \beta \\
= \beta_0 + \beta_1 X_{ij}^{(1)} + \beta_2 X_{ij}^{(2)} + \cdots + \beta_p X_{ij}^{(p)}
\]

- $Y_{ij}$ are assumed to be independent, given $X$’s
- $X_{ij}$ can encode edge covariates, or network statistics that do not violate indep of the $Y_{ij}$’s (basically, dyad statistics)

---

The Simplest Dyadic Independence Model: Erdos-Renyi-Gilbert

- All ties have the same probability
- Can estimate this as the tie density in the graph $P(\text{edge}) = K/(N^*(N-1)) = 32/90 = 0.3555556$

You can also fit Dyadic Independence Model with only an intercept and estimate it.

```r
diag(y) <- NA
> y <- c(y)
> e.r.g <- glm(y ~ 1, 
+  family=binomial)
> coef(e.r.g)
  (Intercept) 
-0.5947071
> exp(-0.5947)/
+ (1 + exp(-0.5947))
[1] 0.3555572
```
A slightly fancier model: The sender-receiver model

- \( \logit(P[Y_{ij}=1]) = \alpha_i + \beta_j \)
  - \( \alpha_i \) is the propensity to send a tie
  - \( \beta_j \) is the propensity to receive a tie

- Setting up an \( X \) matrix to give this model the form \( \logit(P[Y_{ij}=1]) = X\beta \) is a bit of work – see the R notes...

- From the fit on the next page, we see that the only effect different from E-R-G is a receiver effect for BWV!

The Sender-Receiver Model

```r
> # setting up X is in the R notes...
> ab.model <- glm(y ~ X, + family=binomial)
> summary(ab.model)
```

```
Coefficients: (2 not defined because of singularities)

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>z</td>
<td>Est</td>
<td>SE</td>
<td>z</td>
</tr>
<tr>
<td>(Int)</td>
<td>-1.543e+00</td>
<td>1.076e+00</td>
<td>-1.434</td>
<td>Xb.KPJ</td>
<td>5.700e-01</td>
<td>1.079e+00</td>
</tr>
<tr>
<td>Xa.KPJ</td>
<td>-1.049e-15</td>
<td>1.050e+00</td>
<td>0.000</td>
<td>Xb.WQM</td>
<td>-1.419e-15</td>
<td>1.145e+00</td>
</tr>
<tr>
<td>Xa.WQM</td>
<td>-1.531e-16</td>
<td>1.050e+00</td>
<td>0.000</td>
<td>Xb.FOM</td>
<td>1.487e+00</td>
<td>1.048e+00</td>
</tr>
<tr>
<td>Xa.FOM</td>
<td>5.098e-01</td>
<td>1.015e+00</td>
<td>0.502</td>
<td>Xb.SAE</td>
<td>1.045e+00</td>
<td>1.053e+00</td>
</tr>
<tr>
<td>Xa.SAE</td>
<td>-1.532e+00</td>
<td>1.338e+00</td>
<td>-1.145</td>
<td>Xb.NYZ</td>
<td>-2.639e-15</td>
<td>1.145e+00</td>
</tr>
<tr>
<td>Xa.NYZ</td>
<td>5.098e-01</td>
<td>1.015e+00</td>
<td>0.502</td>
<td>Xb.YAW</td>
<td>5.700e-01</td>
<td>1.079e+00</td>
</tr>
<tr>
<td>Xa.YAW</td>
<td>5.098e-01</td>
<td>1.015e+00</td>
<td>0.502</td>
<td>Xb.EVN</td>
<td>-2.059e-15</td>
<td>1.145e+00</td>
</tr>
<tr>
<td>Xa.EVN</td>
<td>9.722e-01</td>
<td>1.004e+00</td>
<td>0.968</td>
<td>Xb.BWV</td>
<td>2.423e+00</td>
<td>1.097e+00</td>
</tr>
<tr>
<td>Xa.BWV</td>
<td>4.638e-16</td>
<td>1.050e+00</td>
<td>0.000</td>
<td>Xb.WAP</td>
<td>-2.351e-15</td>
<td>1.145e+00</td>
</tr>
<tr>
<td>Xa.WAP</td>
<td>-6.213e-01</td>
<td>1.127e+00</td>
<td>-0.551</td>
<td>Xb.REK</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Xa.REK</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>---</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
Covariates in the Pitts-Spillane networks that might affect ties...

- We can explore a **homophily** effect: Is teaching in the same grade associated with greater advice-seeking?
- It seems to depend on the school!
  - **In school 1 it matters!**
  - **In school 11 it does not...**

```r
> Y <- extract.Y(1)  # see the R notes
> X <- extract.X(1)  # for this...
> test01.glm <- glm(Y ~ teach.same.grade,
+ family=binomial, data=as.data.frame(X))
> summary(test01.glm)

Est      SE       Z
(Int)  -2.8536  0.1626 -17.549 ***
same.grade  1.1532  0.2877   4.009 ***
```

```r
> Y <- extract.Y(11)
> X <- extract.X(11)
> test11.glm <- glm(Y ~ teach.same.grade,
+ family=binomial, data=as.data.frame(X))
> summary(test11.glm)

Est         SE      Z
(Int)       -1.9833     0.2754 -7.202 ***
same.grade -15.5828  1398.7210 -0.011
```

---

**Digression to R...**

- For the 15 schools from Pitts & Spillane (2009):
  - The Y’s record ties (adjacency matrix)
  - The X’s are external covariates, as follows:

  **For Dyads:**
  - same.yrs.in.schl
  - same.innov.attitude
  - teach.same.grade

  **For Teachers:**
  - yrs.tchg.sender
  - tchr.trust.sender
  - yrs.tchg.recvr
  - tchr.trus.recvr

  **For Schools:**
  - catholic
  - school.size

- extract.Y(m) extracts the Y’s for school m
- extract.X(m) extracts the X’s for school m

---
Conditionally Independent Dyad (CID) models (Andrew, Next!)

- CID models generalize Dyadic Independence models by adding a latent variable:
  \[ g(E[Y_{ij}]) = X_{ij}\beta + U_{ij} \]
  - \( X_{ij}\beta \) are edge covariates (that preserve independence of dyads)
  - \( U_{ij} \) is a random effect, i.e. latent/unobserved structure
    - Allows for some structured dependence across dyads
      - Clustering, block/community structure, transitivity...
    - \textbf{Still have } \( Y_{ij} \perp Y_{lm} \mid U_{ij}, U_{lm} \text{ whenever } (ij) \neq (lm) \)

Hierarchical Network Models (HNM) (Tracy, later today!)

- Model common elements of an ensemble of networks
- Gain power to detect effects of interventions and other covariates