Assortativity and Mixing
Complex Networks, CSYS/MATH 303, Spring, 2010

Prof. Peter Dodds

Department of Mathematics & Statistics
Center for Complex Systems
Vermont Advanced Computing Center
University of Vermont

Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.
Outline

Definition

General mixing

Assortativity by degree

Contagion

References
Basic idea:

- Random networks with arbitrary degree distributions cover much territory but do not represent all networks.
- Moving away from pure random networks was a key first step.
- We can extend in many other directions and a natural one is to introduce correlations between different kinds of nodes.
- Node attributes may be anything, e.g.:
  1. degree
  2. demographics (age, gender, etc.)
  3. group affiliation
- We speak of mixing patterns, correlations, biases...
- Networks are still random at base but now have more global structure.
- Build on work by Newman [3, 4].
General mixing between node categories

- Assume types of nodes are countable, and are assigned numbers 1, 2, 3, \ldots.
- Consider networks with directed edges.

\[
e_{\mu \nu} = \Pr\left( \text{an edge connects a node of type } \mu \text{ to a node of type } \nu \right)
\]

\[
a_\mu = \Pr(\text{an edge comes from a node of type } \mu)
\]

\[
b_\nu = \Pr(\text{an edge leads to a node of type } \nu)
\]

- Write \( \mathbf{E} = [e_{\mu \nu}] \), \( \vec{a} = [a_\mu] \), and \( \vec{b} = [b_\nu] \).
- Requirements:

\[
\sum_{\mu} e_{\mu \nu} = 1, \quad \sum_{\nu} e_{\mu \nu} = a_\mu, \quad \text{and} \quad \sum_{\mu} e_{\mu \nu} = b_\nu.
\]
Connection to degree distribution:
Notes:

- Varying $e_{\mu\nu}$ allows us to move between the following:

  1. **Perfectly assortative networks** where nodes only connect to like nodes, and the network breaks into subnetworks.
     
     Requires $e_{\mu\nu} = 0$ if $\mu \neq \nu$ and $\sum_{\mu} e_{\mu\mu} = 1$.

  2. **Uncorrelated networks** (as we have studied so far)
     
     For these we must have independence: $e_{\mu\nu} = a_\mu b_\nu$.

  3. **Disassortative networks** where nodes connect to nodes distinct from themselves.

- Disassortative networks can be hard to build and may require constraints on the $e_{\mu\nu}$.

- Basic story: level of assortativity reflects the degree to which nodes are connected to nodes within their group.
Correlation coefficient:

- Quantify the level of assortativity with the following assortativity coefficient \(^4\):

\[
r = \frac{\sum_{\mu} e_{\mu\mu} - \sum_{\mu} a_{\mu} b_{\mu}}{1 - \sum_{\mu} a_{\mu} b_{\mu}} = \frac{\text{Tr} \ E - \| E^2 \|_1}{1 - \| E^2 \|_1}
\]

where \( \| \cdot \|_1 \) is the 1-norm = sum of a matrix’s entries.

- \( \text{Tr} \ E \) is the fraction of edges that are within groups.
- \( \| E^2 \|_1 \) is the fraction of edges that would be within groups if connections were random.
- \( 1 - \| E^2 \|_1 \) is a normalization factor so \( r_{\text{max}} = 1 \).
- When \( \text{Tr} e_{\mu\mu} = 1 \), we have \( r = 1 \).
- When \( e_{\mu\mu} = a_{\mu} b_{\mu} \), we have \( r = 0 \).
Correlation coefficient:

Notes:

- $r = -1$ is inaccessible if three or more types are presents.
- Disassortative networks simply have nodes connected to unlike nodes—no measure of how unlike nodes are.
- Minimum value of $r$ occurs when all links between non-like nodes: $\text{Tr} e_{\mu\mu} = 0$.

$$r_{\text{min}} = \frac{-\|E^2\|_1}{1 - \|E^2\|_1}$$

where $-1 \leq r_{\text{min}} < 0$. 
Scalar quantities

- Now consider nodes defined by a scalar integer quantity.
- Examples: age in years, height in inches, number of friends, ...
- $e_{jk} = \Pr$ a randomly chosen edge connects a node with value $j$ to a node with value $k$.
- $a_j$ and $b_k$ are defined as before.
- Can now measure correlations between nodes based on this scalar quantity using standard Pearson correlation coefficient $\rho$:

$$
r = \frac{\sum_{jk} jk(e_{jk} - a_j b_k)}{\sigma_a \sigma_b} = \frac{\langle jk \rangle - \langle j \rangle_a \langle k \rangle_b}{\sqrt{\langle j^2 \rangle_a - \langle j \rangle_a^2} \sqrt{\langle k^2 \rangle_b - \langle k \rangle_b^2}}
$$

- This is the observed normalized deviation from randomness in the product $jk$. 

$\rho$
Degree-degree correlations

- Natural correlation is between the degrees of connected nodes.
- Now define $e_{jk}$ with a slight twist:

$$e_{jk} = \Pr \left( \begin{array}{c}
\text{an edge connects a degree } j + 1 \text{ node} \\
\text{to a degree } k + 1 \text{ node}
\end{array} \right)$$

$$= \Pr \left( \begin{array}{c}
\text{an edge runs between a node of in-degree } j \\
\text{and a node of out-degree } k
\end{array} \right)$$

- Useful for calculations (as per $R_k$)
- **Important**: Must separately define $P_0$ as the $\{e_{jk}\}$ contain no information about isolated nodes.
- Directed networks still fine but we will assume from here on that $e_{jk} = e_{kj}$. 
Degree-degree correlations

- Notation reconciliation for undirected networks:

\[ r = \frac{\sum_{j} k_j k (e_{jk} - R_j R_k)}{\sigma^2_R} \]

where, as before, \( R_k \) is the probability that a randomly chosen edge leads to a node of degree \( k + 1 \), and

\[ \sigma^2_R = \sum_j j^2 R_j - \left( \sum_j j R_j \right)^2. \]
Degree-degree correlations

Error estimate for $r$:

- Remove edge $i$ and recompute $r$ to obtain $r_i$.
- Repeat for all edges and compute using the jackknife method \(^1\)

$$\sigma_r^2 = \sum_i (r_i - r)^2.$$

- Mildly sneaky as variables need to be independent for us to be truly happy and edges are correlated...
Measurements of degree-degree correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>Network</th>
<th>Type</th>
<th>Size $n$</th>
<th>Assortativity $r$</th>
<th>Error $\sigma_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td>Physics coauthorship</td>
<td>undirected</td>
<td>52 909</td>
<td>0.363</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Biology coauthorship</td>
<td>undirected</td>
<td>1 520 251</td>
<td>0.127</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>Mathematics coauthorship</td>
<td>undirected</td>
<td>253 339</td>
<td>0.120</td>
<td>0.002</td>
</tr>
<tr>
<td>c</td>
<td>Film actor collaborations</td>
<td>undirected</td>
<td>449 913</td>
<td>0.208</td>
<td>0.0002</td>
</tr>
<tr>
<td>d</td>
<td>Company directors</td>
<td>undirected</td>
<td>7 673</td>
<td>0.276</td>
<td>0.004</td>
</tr>
<tr>
<td>e</td>
<td>Student relationships</td>
<td>undirected</td>
<td>573</td>
<td>$-0.029$</td>
<td>0.037</td>
</tr>
<tr>
<td>f</td>
<td>Email address books</td>
<td>directed</td>
<td>16 881</td>
<td>0.092</td>
<td>0.004</td>
</tr>
<tr>
<td>g</td>
<td>Power grid</td>
<td>undirected</td>
<td>4 941</td>
<td>$-0.003$</td>
<td>0.013</td>
</tr>
<tr>
<td>h</td>
<td>Internet</td>
<td>undirected</td>
<td>10 697</td>
<td>$-0.189$</td>
<td>0.002</td>
</tr>
<tr>
<td>i</td>
<td>World Wide Web</td>
<td>directed</td>
<td>269 504</td>
<td>$-0.067$</td>
<td>0.0002</td>
</tr>
<tr>
<td>j</td>
<td>Software dependencies</td>
<td>directed</td>
<td>3 162</td>
<td>$-0.016$</td>
<td>0.020</td>
</tr>
<tr>
<td>k</td>
<td>Protein interactions</td>
<td>undirected</td>
<td>2 115</td>
<td>$-0.156$</td>
<td>0.010</td>
</tr>
<tr>
<td>l</td>
<td>Metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>$-0.240$</td>
<td>0.007</td>
</tr>
<tr>
<td>m</td>
<td>Neural network</td>
<td>directed</td>
<td>307</td>
<td>$-0.226$</td>
<td>0.016</td>
</tr>
<tr>
<td>n</td>
<td>Marine food web</td>
<td>directed</td>
<td>134</td>
<td>$-0.263$</td>
<td>0.037</td>
</tr>
<tr>
<td>o</td>
<td>Freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>$-0.326$</td>
<td>0.031</td>
</tr>
</tbody>
</table>

- Social networks tend to be assortative (homophily)
- Technological and biological networks tend to be disassortative
Next: Generalize our work for random networks to degree-correlated networks.

As before, by allowing that a node of degree $k$ is activated by one neighbor with probability $b_{k1}$, we can handle various problems:

1. find the giant component size.
2. find the probability and extent of spread for simple disease models.
3. find the probability of spreading for simple threshold models.
Goal: Find \( f_{n,j} = \text{Pr} \) an edge emanating from a degree \( j+1 \) node leads to a finite active subcomponent of size \( n \).

Repeat: a node of degree \( k \) is in the game with probability \( b_{k1} \).

Define \( \vec{b}_1 = [b_{k1}] \).

Plan: Find the generating function
\[
F_j(x; \vec{b}_1) = \sum_{n=0}^{\infty} f_{n,j} x^n.
\]
Spreading on degree-correlated networks

- **Recursive relationship:**
  \[
  F_j(x; \vec{b}_1) = x^0 \sum_{k=0}^{\infty} \frac{e_{jk}}{R_j}(1 - b_{k+1,1}) + x \sum_{k=0}^{\infty} \frac{e_{jk}}{R_j} b_{k+1,1} \left[ F_k(x; \vec{b}_1) \right]^k .
  \]

- **First term** = \(\text{Pr}\) that the first node we reach is not in the game.
- **Second term** involves \(\text{Pr}\) we hit an active node which has \(k\) outgoing edges.
- Next: find average size of active components reached by following a link from a degree \(j + 1\) node = \(F'_j(1; \vec{b}_1)\).
Spreading on degree-correlated networks

- Differentiate $F_j(x; \vec{b}_1)$, set $x = 1$, and rearrange.
- We use $F_k(1; \vec{b}_1) = 1$ which is true when no giant component exists. We find:

$$R_j F_j'(1; \vec{b}_1) = \sum_{k=0}^{\infty} e_{jk} b_{k+1,1} + \sum_{k=0}^{\infty} k e_{jk} b_{k+1,1} F_k'(1; \vec{b}_1).$$

- Rearranging and introducing a sneaky $\delta_{jk}$:

$$\sum_{k=0}^{\infty} \left( \delta_{jk} R_k - k b_{k+1,1} e_{jk} \right) F_k'(1; \vec{b}_1) = \sum_{k=0}^{\infty} e_{jk} b_{k+1,1}. $$
Spreading on degree-correlated networks

- In matrix form, we have

\[ \mathbf{A}_{E, \vec{b}_1} \vec{F}'(1; \vec{b}_1) = \mathbf{E}\vec{b}_1 \]

where

\[
\begin{bmatrix}
A_{E, \vec{b}_1}
\end{bmatrix}_{j+1,k+1} = \delta_{jk} R_k - kb_{k+1,1} e_{jk},
\]

\[
\begin{bmatrix}
\vec{F}'(1; \vec{b}_1)
\end{bmatrix}_{k+1} = F'_k(1; \vec{b}_1),
\]

\[
[\mathbf{E}]_{j+1,k+1} = e_{jk}, \text{ and } \begin{bmatrix}
\vec{b}_1
\end{bmatrix}_{k+1} = b_{k+1,1}.
\]
Spreading on degree-correlated networks

- So, in principle at least:

\[ \bar{F}'(1; \bar{b}_1) = A_{\bar{b}_1}^{-1} \mathbf{E} \bar{b}_1. \]

- Now: as \( \bar{F}'(1; \bar{b}_1) \), the average size of an active component reached along an edge, increases, we move towards a transition to a giant component.

- Right at the transition, the average component size explodes.

- Exploding inverses of matrices occur when their determinants are 0.

- The condition is therefore:

\[ \det A_{\mathbf{E}, \bar{b}_1} = 0 \]
Spreading on degree-correlated networks

- General condition details:

\[ \det A_{E,\vec{b}_1} = \det [\delta_{jk} R_{k-1} - (k - 1)b_{k,1} e_{j-1,k-1}] = 0. \]

- The above collapses to our standard contagion condition when \( e_{jk} = R_j R_k \).

- When \( \vec{b}_1 = \vec{b} \), we have the condition for a simple disease model’s successful spread:

\[ \det [\delta_{jk} R_{k-1} - b(k - 1)e_{j-1,k-1}] = 0. \]

- When \( \vec{b}_1 = \vec{1} \), we have the condition for the existence of a giant component:

\[ \det [\delta_{jk} R_{k-1} - (k - 1)e_{j-1,k-1}] = 0. \]

- Bonusville: We’ll find another (possibly better) version of this set of conditions later...
We’ll next find two more pieces:

1. $P_{\text{trig}}$, the probability of starting a cascade
2. $S$, the expected extent of activation given a small seed.

**Triggering probability:**

- Generating function:

$$H(x; \vec{b}_1) = x \sum_{k=0}^{\infty} P_k \left[ F_{k-1}(x; \vec{b}_1) \right]^k.$$ 

- Generating function for vulnerable component size is more complicated.
Spreading on degree-correlated networks

- Want probability of **not reaching** a finite component.

\[ P_{\text{trig}} = S_{\text{trig}} = 1 - H(1; \vec{b}_1) = 1 - \sum_{k=0}^{\infty} P_k \left[ F_{k-1}(1; \vec{b}_1) \right]^k. \]

- Last piece: we have to compute \( F_{k-1}(1; \vec{b}_1) \).

- Nastier (nonlinear)—we have to solve the recursive expression we started with when \( x = 1 \):

\[ F_j(1; \vec{b}_1) = \sum_{k=0}^{\infty} \frac{e_{jk}}{R_j} (1 - b_{k+1, 1}) + \sum_{k=0}^{\infty} \frac{e_{jk}}{R_j} b_{k+1, 1} \left[ F_k(1; \vec{b}_1) \right]^k. \]

- Iterative methods should work here.
Spreading on degree-correlated networks

- **Truly final piece:** Find final size using approach of Gleeson [2], a generalization of that used for uncorrelated random networks.
- Need to compute $\theta_{j,t}$, the probability that an edge leading to a degree $j$ node is infected at time $t$.
- Evolution of edge activity probability:

$$
\theta_{j,t+1} = G_j(\vec{\theta}_t) = \phi_0 + (1 - \phi_0) \times \\
\sum_{k=1}^{\infty} \frac{e_{j-1,k-1}}{R_{j-1}} \sum_{i=0}^{k-1} \binom{k-1}{i} \theta_{k,t} (1 - \theta_{k,t})^{k-1-i} b_{ki}.
$$

- Overall active fraction’s evolution:

$$
\phi_{t+1} = \phi_0 + (1 - \phi_0) \sum_{k=0}^{\infty} P_k \sum_{i=0}^{k} \binom{k}{i} \theta_{k,t} (1 - \theta_{k,t})^{k-i} b_{ki}.
$$
Spreading on degree-correlated networks

- As before, these equations give the actual evolution of $\phi_t$ for synchronous updates.
- Contagion condition follows from $\vec{\theta}_{t+1} = \vec{G}(\vec{\theta}_t)$.
- Expand $\vec{G}$ around $\vec{\theta}_0 = \vec{0}$.

$$\theta_{j,t+1} = G_j(\vec{0}) + \sum_{k=1}^{\infty} \frac{\partial G_j(\vec{0})}{\partial \theta_k,t} \theta_{k,t} + \sum_{k=1}^{\infty} \frac{\partial^2 G_j(\vec{0})}{\partial \theta^2_k,t} \theta_{k,t}^2 + \ldots$$

- If $G_j(\vec{0}) \neq 0$ for at least one $j$, always have some infection.
- If $G_j(\vec{0}) = 0 \ \forall \ j$, largest eigenvalue of $\left[ \frac{\partial G_j(\vec{0})}{\partial \theta_k,t} \right]$ must exceed 1.
- Condition for spreading is therefore dependent on eigenvalues of this matrix:

$$\frac{\partial G_j(\vec{0})}{\partial \theta_k,t} = \frac{e_{j-1,k-1}}{R_{j-1}} (k - 1) b_{k1}$$

Insert question from assignment 6 (⊞)
How the giant component changes with assortativity

![Graph showing how the giant component changes with assortativity](https://example.com/graph)

- More assortative networks percolate for lower average degrees
- But disassortative networks end up with higher extents of spreading.

from Newman, 2002 [3]
References

The jackknife estimate of variance.

Cascades on correlated and modular random networks.

Assortative mixing in networks.

Mixing patterns in networks.