Social Network Analysis

Some content from Lada Adamic and Eytan Adar
Figure 1
Vocabulary Lesson

**Actor**
- Person
- Group
- Event
- ...

**Relational Tie**
- parentOf
- supervisorOf
- reallyHates (+/-)
- ...

**Dyad**

*Relation*: collection of ties of a specific type (every parentOf tie)
Vocabulary Lesson

If A likes B and B likes C then A likes C (transitivity)
If A likes B and C likes B then A likes C

...
Vocabulary Lesson

Social Network

One mode
Vocabulary Lesson

Social Network

Two mode
Vocabulary Lesson

Ego-Centered Network
(egonet, neighborhood)
Describing Networks

- **Geodesic**
  - $\text{shortest\_path}(n,m)$

- **Diameter**
  - $\text{max}(\text{geodesic}(n,m))$ n,m actors in graph

- **Density / Sparsity**
  - Number of existing edges / All possible edges
  - Degeneracy (number $k$ such that every subgraph has a vertex of degree $k$ or less)
  - Related to arboricity (number of forests that cover every edge)
### Degeneracy in the Real World

<table>
<thead>
<tr>
<th>graph</th>
<th>n</th>
<th>m</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>zacahry [48]</td>
<td>34</td>
<td>78</td>
<td>4</td>
</tr>
<tr>
<td>dolphins [35]</td>
<td>62</td>
<td>159</td>
<td>4</td>
</tr>
<tr>
<td>power [47]</td>
<td>4,941</td>
<td>6,594</td>
<td>5</td>
</tr>
<tr>
<td>polbooks [28]</td>
<td>105</td>
<td>441</td>
<td>6</td>
</tr>
<tr>
<td>adjnoun [29]</td>
<td>112</td>
<td>425</td>
<td>6</td>
</tr>
<tr>
<td>football [15]</td>
<td>115</td>
<td>613</td>
<td>8</td>
</tr>
<tr>
<td>lmesis [25]</td>
<td>77</td>
<td>254</td>
<td>9</td>
</tr>
<tr>
<td>celegensneural [47]</td>
<td>297</td>
<td>1,248</td>
<td>9</td>
</tr>
<tr>
<td>netscience [39]</td>
<td>1,589</td>
<td>2,742</td>
<td>19</td>
</tr>
<tr>
<td>internet [40]</td>
<td>22,963</td>
<td>48,421</td>
<td>25</td>
</tr>
<tr>
<td>condmat-2005 [38]</td>
<td>40,421</td>
<td>175,693</td>
<td>29</td>
</tr>
<tr>
<td>polblogs [4]</td>
<td>1,490</td>
<td>16,715</td>
<td>36</td>
</tr>
<tr>
<td>astro-ph [38]</td>
<td>16,706</td>
<td>121,251</td>
<td>56</td>
</tr>
</tbody>
</table>

From https://arxiv.org/pdf/1006.5440
Degeneracy in the Real World

<table>
<thead>
<tr>
<th>graph</th>
<th>n</th>
<th>m</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>1,455</td>
<td>1,636</td>
<td>6</td>
</tr>
<tr>
<td>worm</td>
<td>3,518</td>
<td>3,518</td>
<td>10</td>
</tr>
<tr>
<td>plant</td>
<td>1,745</td>
<td>3,098</td>
<td>12</td>
</tr>
<tr>
<td>fruitfly</td>
<td>7,282</td>
<td>24,894</td>
<td>12</td>
</tr>
<tr>
<td>human</td>
<td>9,527</td>
<td>31,182</td>
<td>12</td>
</tr>
<tr>
<td>fission-yeast</td>
<td>2,031</td>
<td>12,637</td>
<td>34</td>
</tr>
<tr>
<td>yeast</td>
<td>6,008</td>
<td>156,945</td>
<td>64</td>
</tr>
</tbody>
</table>
Degeneracy in the Real World

<table>
<thead>
<tr>
<th>graph</th>
<th>n</th>
<th>m</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>roadNet-CA [34]</td>
<td>1,965,206</td>
<td>2,766,607</td>
<td>3</td>
</tr>
<tr>
<td>roadNet-PA [34]</td>
<td>1,088,092</td>
<td>1,541,898</td>
<td>3</td>
</tr>
<tr>
<td>roadNet-TX [34]</td>
<td>1,379,917</td>
<td>1,921,660</td>
<td>3</td>
</tr>
<tr>
<td>amazon0601 [30]</td>
<td>403,394</td>
<td>2,443,408</td>
<td>10</td>
</tr>
<tr>
<td>email-EuAll [31]</td>
<td>265,214</td>
<td>364,481</td>
<td>37</td>
</tr>
<tr>
<td>email-Enron [24]</td>
<td>36,692</td>
<td>183,831</td>
<td>43</td>
</tr>
<tr>
<td>web-Google [2]</td>
<td>875,713</td>
<td>4,322,051</td>
<td>44</td>
</tr>
<tr>
<td>soc-wiki-Vote [33]</td>
<td>7,115</td>
<td>100,762</td>
<td>53</td>
</tr>
<tr>
<td>soc-slashdot0902 [34]</td>
<td>82,168</td>
<td>504,230</td>
<td>55</td>
</tr>
<tr>
<td>cit-Patents [18]</td>
<td>3,774,768</td>
<td>16,518,947</td>
<td>64</td>
</tr>
<tr>
<td>soc-Epinions1 [42]</td>
<td>75,888</td>
<td>405,740</td>
<td>67</td>
</tr>
<tr>
<td>soc-wiki-Talk [33]</td>
<td>2,394,385</td>
<td>4,659,565</td>
<td>131</td>
</tr>
<tr>
<td>web-berkstan [34]</td>
<td>685,231</td>
<td>6,649,470</td>
<td>201</td>
</tr>
</tbody>
</table>

From https://arxiv.org/pdf/1006.5440
Random Network Graph Models

• Two classic examples:
  – Erdős–Rényi
    • $G(n,M)$: randomly draw $M$ edges between $n$ nodes
    • $G(n,p)$: randomly draw edges between $n$ nodes, each with probability $p$.
  – These models don’t really model the real world, in that they don’t show:
    • Small world phenomenon
    • Power laws
    • Sparsity
Milgram’s experiment (1960’s):
- Given a target individual and a particular property, pass the message to a person you correspond with who is “closest” to the target.
- “Six degrees of separation”
Two more examples of power laws

Distribution of users among web sites

Sites ranked by popularity
Power Laws (Scale-Free Networks)

• Power-law
  – A scale-free network is a network whose degree distribution follows a power law, at least asymptotically.
  – That is, the fraction \( P(k) \) of nodes in the network having \( k \) connections to other nodes goes for large values of \( k \) as
    \[
    P(k) \sim x^{-k}
    \]
  – Typically \( k \) is in the range from 2 to 3.
  – Many networks have been reported to be scale-free.
Barabási & Albert (BA) Random Graph Model

- Very simple algorithm to implement
  - start with an initial set of $m_0$ fully connected nodes
    - e.g. $m_0 = 3$
  - now add new vertices one by one, each one with exactly $m$ edges
  - each new edge connects to an existing vertex in proportion to the number of edges that vertex already has → preferential attachment
Properties of a BA graph

• The degree distribution is scale free with exponent $k = 3$  
  $P(k) = 2 m^2 / k^3$

• The graph is connected
  – Every new vertex is born with a link or several links. It then connects to $m$ ‘older’ vertices
  – Probability $p_i$ of connecting to node $i$:
    • $k_i$ is the degree of node $i$

• The older get richer
  – Nodes accumulate links as time goes on, which gives older nodes an advantage since newer nodes are going to attach preferentially – and older nodes have a higher degree to tempt them with than some new kid on the block

\[ p_i = \frac{k_i}{\sum_j k_j} \]
Common Tasks

• Measuring “importance”
  – Centrality, prestige
• Diffusion modeling
  – Epidemiological
• Clustering
  – Clustering coefficients
• Structure analysis
  – Subgraph isomorphisms, etc.
• Visualization/Privacy/etc.
Centrality Measures

• Degree centrality
  – Edges per node (the more, the more important the node)

• Closeness centrality
  – How close the node is to every other node

• Betweenness centrality
  – How many shortest paths go through the edge node (communication metaphor)
Common Tasks

• Measuring “importance”
  – Centrality, prestige (incoming links)

• **Diffusion modeling**
  – Epidemiological

• Clustering
  – Clustering coefficients

• Structure analysis
  – Subgraph Isomorphisms, etc.

• Visualization/Privacy/etc.
Epidemiological

- Viruses
  - Biological, computational
  - STDs, needle sharing, etc.
  - Mark Handcock at UW

- Blog networks
  - Applying SIR models (Info Diffusion Through Blogspace, Gruhl et al.)
    - Induce transmission graph, cascade models, simulation
  - Link prediction (Tracking Information Epidemics in Blogspace, Adar et al.)
    - Find repeated “likely” infections
  - Outbreak detection (Cost-effective Outbreak Detection in Networks, Leskovec et al.)
    - Submodularity
Common Tasks

• Measuring “importance”
  – Centrality, prestige (incoming links)
• Diffusion modeling
  – Epidemiological
• Clustering
  – Clustering coefficients
• Structure analysis
  – Subgraph Isomorphisms, etc.
• Visualization/Privacy/etc.
Blockmodel of U.S. Philosophy Departments. Note that row/column numbers do not correspond to PGR rankings.
<table>
<thead>
<tr>
<th>Domingo</th>
<th>Carlos</th>
<th>Alejandro</th>
<th>Eduardo</th>
<th>Frank</th>
<th>Hal</th>
<th>Karl</th>
<th>Bob</th>
<th>Ike</th>
<th>Gill</th>
<th>Lanny</th>
<th>Mike</th>
<th>John</th>
<th>Xavier</th>
<th>Utrecht</th>
<th>Norm</th>
<th>Russ</th>
<th>Quint</th>
<th>Wendle</th>
<th>Ozzie</th>
<th>Ted</th>
<th>Sam</th>
<th>Vern</th>
<th>Paul</th>
</tr>
</thead>
</table>
Global Clustering Coefficient

- The global clustering coefficient $C$ is defined as:

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triplets of vertices}}.$$ 

- In this formula, a connected triplet is defined to be a connected subgraph consisting of three vertices and two edges. Thus, each triangle forms three connected triplets, explaining the factor of three in the formula.
Local Clustering Coefficient

- The local clustering coefficient of a vertex (node) in a graph quantifies how close its neighbors are to being a clique (i.e., complete graph).
- The number of possible connections for the neighbors of a node $i$ of degree $k_i$ is, of course, $k_i(k_i - 1)/2$.
- The local clustering coefficient $C_i$ of node $i$ is defined as:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

- We will discuss later how to compute these values.
Common Tasks

• Measuring “importance”
  – Centrality, prestige (incoming links)

• Diffusion modeling
  – Epidemiological

• Clustering
  – Blockmodeling, Girvan-Newman

• Structure analysis
  – Subgraph Isomorphisms, etc.

• Visualization/Privacy/etc.
Common Tasks

• Measuring “importance”
  – Centrality, prestige (incoming links)
• Diffusion modeling
  – Epidemiological
• Clustering
  – Clustering coefficients
• Structure analysis
  – Motifs, Isomorphisms, etc.
• Visualization/Privacy/etc.
Privacy

• Emerging interest in anonymizing networks
  – Lars Backstrom (WWW’07) demonstrated one of the first attacks

• How to remove labels while preserving graph properties?
  – While ensuring that labels cannot be reapplied