What does a real-world network look like?

Reading: Ch 2 of EK, Ch 2 & 3 of Jackson
Graph visualization using Gephi

Mohammad T. Irfan
Email: mirfan@bowdoin.edu

Real-World Networks
And their common properties
1. Macro-level (graph-level)
2. Micro-level (node-level)
Macro-level properties

1. Giant component
2. Small-world
3. Degree distribution
4. Clustering

1. Giant component

- Intuitive example: world acquaintance network
- Questions
  - Is it connected?
  - How many giant components are there?
Examples

- Actor network
  - Edge between two actors iff they appear together in a movie
  - 98% of 449,913 actors belong to the giant component (IMDB, May 2000)

More examples

- Instant messaging
- Co-author network
- Email
- Biological networks (neural networks)
- Technology networks (power grid)
- The Internet (web of links)

Can you think of a network that doesn’t have any giant component?
What is the implication?

High school relationships (1993-95)

2. Small-world property

- Proposition
  - The average shortest path between any two nodes in a connected component is “small”

- Intuition

Also known as distance
Six degrees of separation

- Hungarian author Frigyes Karinthy (1929 short story “Chain-Links”)
  “A fascinating game grew out of this discussion. One of us suggested performing the following experiment to prove that the population of the Earth is closer together now than they have ever been before. We should select any person from the 1.5 billion inhabitants of the Earth - anyone, anywhere at all. He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances.”

- John Guare’s play (1990) & later movie

Milgram’s experiment (1963)
Milgram’s experiment (cont...)

Critiques

- Only 64 out of 296 cases were successful
- How useful? What is the implication?
  - Milgram: “six worlds apart”
Contagion of TB
(Valdis Krebs, Oklahoma, 2002)

Another example

- Microsoft instant messenger (2008)
  - 240M node network
  - Edge: Two-way conversation at some point during a month-long observation period
  - Average distance: 6.6

Fraction of pairs of nodes having this distance
Computational question

- How to find the “right 6 people?”
  - Breadth-first search (BFS) algorithm to find the shortest path
- Fun application- Bacon number
  - Bacon number of an actor = distance from Kevin Bacon
  - Average Bacon number: 2.9
  - [https://oracleofbacon.org/](https://oracleofbacon.org/)

Shortest path algorithm

Breadth-First Search (BFS)
**BFS algorithm**

- **Resulting graph: BFS tree**

  - Ye u
  - Distance = 0
  - Distance = 1
  - Distance = 2
  - Distance = 3

  - AKA "root"
  - Other existing edges within a layer are not drawn here. Draw only the edges explored.
  - Your friends
  - Friends of friends
  - Friends of friends of friends

  - Nodes whose distance have not yet been calculated and who have edges to nodes in the previous layer

**Exercise: Draw BFS from MIT**

ARPANET (1970)
When does BFS give shortest paths?

- When all the edges have the same "weight"/dist.
- Negative example:
Some special types of graphs

- **Tree**
  - Connected, acyclic graph
  - Example: BFS tree

- **Bipartite graph**
  - Two sets of nodes with no edge within the same set of nodes
  - Example: Network between movies and actors

3. Degree distribution

- What’s the probability of finding a node with degree k?
- What fraction of nodes have degree k? Call it $P_k$. 
Real-world degree distributions

- Power law distribution (or Pareto distrib.) vs. normal distribution
- Mathematical formulation
- Scale-free networks

4. Clustering coefficients

- Clustering coeff = Average probability that two friends of a node are also friends
- How to calculate?
  - Local clustering coeff. of node i,
    \[ C_i = \frac{\text{Actual # of edges among i's friends}}{\text{Max possible # of edges among i's friends}} \]
  - Clustering coefficient of the whole network
    = average \( C_i \) of all the nodes i

Need to count

\[ d_i \cdot (d_i - 1) / 2 \] where \( d_i = \text{degree of } i \]
Example

What is the clustering coefficient of this network?

Political blogs (2004)

“High” clustering coefficient is observed in real-world networks
Empirical study of network properties

- Uzzi et al., 2007

N = # of nodes
k = Avg degree
L = Avg shortest path length
CC = Clustering coefficient

<table>
<thead>
<tr>
<th>Authors</th>
<th>Network</th>
<th>Period</th>
<th>N</th>
<th>k</th>
<th>L</th>
<th>CC</th>
<th>Le</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vogel and Walker (2001)</td>
<td>German firms</td>
<td>1970–1997</td>
<td>291</td>
<td>2.02</td>
<td>5.64</td>
<td>3.01</td>
<td>0.84</td>
<td>0.022</td>
</tr>
<tr>
<td>Baron et al. (2005)</td>
<td>Canadian I-beams</td>
<td>1952–1997</td>
<td>55</td>
<td>1.30</td>
<td>3.21</td>
<td>4.50</td>
<td>0.023</td>
<td>0.027</td>
</tr>
<tr>
<td>1969–1974</td>
<td>61</td>
<td>2.22</td>
<td>2.82</td>
<td>3.77</td>
<td>0.283</td>
<td>0.054</td>
<td>0.89</td>
<td>5.24</td>
</tr>
<tr>
<td>1980–1990</td>
<td>142</td>
<td>3.83</td>
<td>2.95</td>
<td>3.16</td>
<td>0.273</td>
<td>0.027</td>
<td>0.94</td>
<td>10.11</td>
</tr>
<tr>
<td>Davis et al. (2003)</td>
<td>US Co. interlocks</td>
<td>1942</td>
<td>135</td>
<td>6.8</td>
<td>3.15</td>
<td>2.7</td>
<td>0.24</td>
<td>0.009</td>
</tr>
<tr>
<td>1999</td>
<td>105</td>
<td>7.2</td>
<td>2.98</td>
<td>2.64</td>
<td>0.2</td>
<td>0.009</td>
<td>3.13</td>
<td>5.13</td>
</tr>
<tr>
<td>Veugelers and Depuydt (2004)</td>
<td>Strategic alliances*</td>
<td>1980–1996</td>
<td>7084</td>
<td>5.29</td>
<td>4.2</td>
<td>3.25</td>
<td>0.34</td>
<td>0.008</td>
</tr>
<tr>
<td>Schilling and Phelps, (forfeiting)</td>
<td>US alliances in 11 2-digit SIC codes**</td>
<td>1992–2000</td>
<td>371</td>
<td>5.11</td>
<td>20.59</td>
<td>3.62</td>
<td>0.26</td>
<td>0.04</td>
</tr>
<tr>
<td>(157)</td>
<td>(1.42)</td>
<td>(18.69)</td>
<td>(3.80)</td>
<td>(0.18)</td>
<td>(0.039)</td>
<td>(2.84)</td>
<td>(7.53)</td>
<td>(2.65)</td>
</tr>
<tr>
<td>Davis et al. (2005)</td>
<td>US Dironco interlocks</td>
<td>1982</td>
<td>2553</td>
<td>19.1</td>
<td>4.55</td>
<td>2.01</td>
<td>0.51</td>
<td>0.009</td>
</tr>
<tr>
<td>1999</td>
<td>2078</td>
<td>17.6</td>
<td>3.98</td>
<td>2.05</td>
<td>0.89</td>
<td>0.009</td>
<td>3.50</td>
<td>98.89</td>
</tr>
<tr>
<td>1999</td>
<td>2356</td>
<td>16.8</td>
<td>3.86</td>
<td>2.09</td>
<td>0.89</td>
<td>0.009</td>
<td>3.53</td>
<td>97.78</td>
</tr>
<tr>
<td>Hering et al. (forthcoming)</td>
<td>US patenting inventors***</td>
<td>1990–1999</td>
<td>7089</td>
<td>4.73</td>
<td>2.73</td>
<td>3.14</td>
<td>0.376</td>
<td>0.005</td>
</tr>
<tr>
<td>(German Co. ownership)</td>
<td>Biology</td>
<td>1993–1997</td>
<td>429</td>
<td>3.56</td>
<td>6.09</td>
<td>5.16</td>
<td>0.63</td>
<td>0.008</td>
</tr>
<tr>
<td>1990–1999</td>
<td>1,520,231</td>
<td>18.1</td>
<td>6.6</td>
<td>0.066</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990–1999</td>
<td>52,909</td>
<td>9.7</td>
<td>5.9</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990–2006</td>
<td>253,339</td>
<td>3.9</td>
<td>7.6</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mosby, 2004</td>
<td>Economics</td>
<td>1965–1999</td>
<td>128,151</td>
<td>9.81</td>
<td>7.57</td>
<td>0.194</td>
<td>0.207</td>
<td>1.30</td>
</tr>
<tr>
<td>1965–1999</td>
<td>87,731</td>
<td>11.53</td>
<td>8.24</td>
<td>0.286</td>
<td>0.042</td>
<td>1.40</td>
<td>0.83</td>
<td>0.61</td>
</tr>
<tr>
<td>1965–1999</td>
<td>68,606</td>
<td>1.241</td>
<td>0.182</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watts (1999)</td>
<td>Film actors</td>
<td>1980–1997</td>
<td>236,880</td>
<td>61</td>
<td>3.65</td>
<td>2.99</td>
<td>0.79</td>
<td>0.00027</td>
</tr>
<tr>
<td>Smith (2006)</td>
<td>U.S. rappeons</td>
<td>2006</td>
<td>3533</td>
<td>3.9</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1279</td>
<td>2.79</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3841</td>
<td>2.3</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watts (2010)</td>
<td>Power grids</td>
<td>1997</td>
<td>4942</td>
<td>2.94</td>
<td>10.7</td>
<td>12.4</td>
<td>0.08</td>
<td>0.005</td>
</tr>
<tr>
<td>1998</td>
<td>3029</td>
<td>3.6</td>
<td>3.8</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>3084</td>
<td>3.6</td>
<td>3.7</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Graph Visualization

Gephi

Links

- Download
  - https://gephi.org/
  - You also need Java 6: https://support.apple.com/kb/dl1572?locale=en_US
Ranking: rank and color nodes, edges, and their labels by numeric properties

Partition: partition nodes and edges

Data Laboratory: Manipulate the input graph files (e.g., apply labels to nodes)

Statistics: Computes graph-level properties. Some of them (e.g., Average Degree) must be done before using other features

Preview: Produces a nice visualization (next slide)

Filters:
- Filter out nodes/edges based on their properties
- Useful filters: Topology

Layout: Select a graph drawing algorithm

Magnifying glass: Centers the graphics

T: Toggle showing node labels

T: Toggle showing edge labels

Slider: Tune the size of node labels

Slider: Tune edge thickness

Useful filters:
- Degree Range

Network among the characters of Les Miserables

Show Labels: Turn it on!

Refresh: Must click this button! Otherwise, nothing will be shown.

To save the visualization as a pdf file:
File → Save
Gephi Vocabulary

<table>
<thead>
<tr>
<th>Term</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>betweenness centrality of a node</td>
<td>how often the node appears on the shortest path between nodes in the network</td>
</tr>
<tr>
<td>closeness centrality of a node</td>
<td>average distance from that node to all other nodes in the network</td>
</tr>
<tr>
<td>degree of a node</td>
<td>the number of edges connected to the node (also connectedness); in a directed graph a node can have in-degree and out-degree measures</td>
</tr>
<tr>
<td>diameter of a graph</td>
<td>the longest shortest path between any two nodes in the graph</td>
</tr>
<tr>
<td>directed graph</td>
<td>this means relationships occur one way only (I follow you, but you do not follow me on Twitter); opposite of undirected (we are friends with each other on Facebook)</td>
</tr>
<tr>
<td>eccentricity of a node</td>
<td>the distance (shortest-path length) from the node to the farthest node from it in the network</td>
</tr>
<tr>
<td>edge</td>
<td>a representation of the connection between two nodes, expresses a relationship (a line)</td>
</tr>
<tr>
<td>eigenvector centrality of a node</td>
<td>in social network analysis, a measure of influence (a node is very influential if it is connected to other influential nodes)</td>
</tr>
<tr>
<td>layout algorithms</td>
<td>also known as graph drawing algorithm; e.g., force-directed drawing where linked nodes attract and non-linked nodes repel</td>
</tr>
<tr>
<td>leaf node</td>
<td>node with a single edge in a “tree-structured” graph</td>
</tr>
<tr>
<td>modularity</td>
<td>a measure of connectedness among groups of nodes (greater than 0.4 is usually considered meaningful)</td>
</tr>
<tr>
<td>node</td>
<td>also called a vertex by mathematicians; a person in a social network graph (a dot or bubble)</td>
</tr>
<tr>
<td>distance from one node to another</td>
<td>the length of the shortest path (counted in the number of edges) from one node to another</td>
</tr>
<tr>
<td>path length</td>
<td>the number of edges in a path</td>
</tr>
<tr>
<td>singleton node or isolated node</td>
<td>node with no edge/connection</td>
</tr>
</tbody>
</table>

Micro-level properties
Centrality
Notation: n = # of nodes
Reading: Jackson (Ch 2)
Centrality measures

1. Degree centrality
2. Closeness centrality
3. Betweenness centrality
4. Prestige/eigenvector centrality

1. Degree centrality
   - A node’s centrality = The node’s degree / (n-1)
   - Who is the most central here?

   ![Diagram of network with nodes and connections]

   1. How about node 4 in this network?
2. Closeness centrality

- Idea: node $i$ is very central if it’s pretty close to the other nodes.
- Avg distance from node $i$ to all other nodes =
  \[
  \frac{\sum_{j \neq i} \text{dist}(i,j)}{n-1}
  \]
- Closeness centrality of $i = 1/\text{Avg distance from } i$
- Example: Compute the closeness centralities of nodes 1 and 4.

3. Betweenness centrality

- Idea: a node $i$ is very central if a lot of the shortest paths go through $i$.
- Betweenness centrality of $i$,
  \[
  \beta_i = \sum_{j,k} \frac{\text{# of shortest paths between } j \text{ and } k \text{ passing through } i}{\text{# of shortest paths between } j \text{ and } k, \text{ irrespective of passing through } i}
  \]
- Exercise: Compute the between centrality of nodes 1, 2, and 3.
  - $\beta_1 = 0$
  - $\beta_2 = 0$
  - $\beta_3 = ?$
- Florentine marriage: Medici most central
Demo

- Gephi

Dolphin network data:
Social network (by association) among 62 dolphins in Doubtful Sound, New Zealand

Demo

- Communities in a network
- Betweenness centrality
  - Gateways to communities
- Demo: Dolphin network
1. Use the modularity statistics with parameter value = 3
2. Partition the nodes by their **Modularity Class** (or community)

4. Compute betweenness centrality (under **Network Diameter**).
5. Rank the nodes by betweenness centrality.

See anything interesting?
Comparison

- What are the differences among:
  - Degree centrality
  - Closeness centrality
  - Betweenness centrality

Matrix algebra

- 4x1 matrix (AKA vector)
  \[
  \begin{bmatrix}
  4 \\
  2.6 \\
  -8.1 \\
  7 \\
  \end{bmatrix}
  \]
- 3x3 matrix
  \[
  \begin{pmatrix}
  1 & 2 & 3 \\
  8 & 4 & 5 \\
  4 & -2 & 6 \\
  \end{pmatrix}
  \]
Matrix multiplication

- $2 \times 3$ matrix multiplied by $3 \times 2$ matrix must match
- Result is a $2 \times 2$ matrix

\[
\begin{pmatrix}
a & b & c \\
d & e & f
\end{pmatrix}
\begin{pmatrix}
u & v \\
w & x \\
y & z
\end{pmatrix} =
\begin{pmatrix}
au + bw + cy & av + bx + cz \\
du + ew + fy & dv + ex + fz
\end{pmatrix}
\]

Transpose of matrix

- Transpose operator: superscript $T$

\[
A =
\begin{pmatrix}
1 & 4 \\
2 & 5 \\
3 & 6
\end{pmatrix}
\]

\[
A^T =
\begin{pmatrix}
1 & 2 & 3 \\
4 & 5 & 6
\end{pmatrix}
\]

- $(A B)^T = B^T A^T$
4. Prestige/Eigenvector/power centrality

- Idea (Phillip Bonacich, 1987): A node’s importance is determined by its friends’ importance
- Mathematical formulation and example

More on eigenvector centrality

- Tutorial on eigenvector
  - Jackson’s Section 2.4 (Appendix)
- Calculating eigenvector
Comparison of centrality measures

A: Degree
B: Closeness
C: Betweenness
D: Eigenvector