Frontiers of Network Science
Fall 2022

Class 5: Random Networks
(Chapter 3 in Textbook)

Boleslaw Szymanski

based on slides by Albert-László Barabási & Roberta Sinatra
CONNECTEDNESS
Connected (undirected) graph: any two vertices can be joined by a path. A disconnected graph is made up by two or more connected components.

Bridge: if we erase it, the graph becomes disconnected.
The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

\[
\begin{pmatrix}
  0 & 1 & 1 & 0 & 0 & 0 & 0 \\
  1 & 0 & 1 & 0 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 1 & 1 \\
  0 & 0 & 0 & 0 & 1 & 1 & 1 \\
  0 & 0 & 0 & 1 & 1 & 0 & 0 \\
  0 & 0 & 0 & 1 & 1 & 0 & 0
\end{pmatrix}
\]

\[
\begin{pmatrix}
  0 & 1 & 1 & 0 & 0 & 0 & 0 \\
  1 & 0 & 1 & 1 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 & 0 & 0 \\
  0 & 1 & 1 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 1 & 1 & 1 \\
  0 & 0 & 0 & 1 & 1 & 0 & 0 \\
  0 & 0 & 0 & 1 & 1 & 0 & 0
\end{pmatrix}
\]
Connectivity of Directed Graphs

Strongly connected directed graph: has a path from each node to every other node and vice versa (e.g. AB path and BA path).

Weakly connected directed graph: it is connected if we disregard the edge directions.

Strongly connected components can be identified, but not every node is part of a nontrivial strongly connected component.

In-component: nodes that can reach the scc,
Out-component: nodes that can be reached from the scc.
Finding the Connected Components of a Network

1. Start from a randomly chosen node $i$ and perform a BFS (BOX 2.5). Label all nodes reached this way with $n = 1$.

2. If the total number of labeled nodes equals $N$, then the network is connected. If the number of labeled nodes is smaller than $N$, the network consists of several components. To identify them, proceed to step 3.

3. Increase the label $n \rightarrow n + 1$. Choose an unmarked node $j$, label it with $n$. Use BFS to find all nodes reachable from $j$, label them all with $n$. Return to step 2.
Clustering coefficient
Clustering coefficient:

what fraction of your neighbors are connected?

Node i with degree $k_i$

$C_i$ in [0,1]

\[
C_i = \frac{2e_i}{k_i(k_i - 1)}
\]

$C_i = 1$

$C_i = 1/2$

$C_i = 0$

Clustering coefficient and Global clustering coefficient

What fraction of your neighbors are connected?

Node i with degree $k_i$

$C_i$ in [0,1]

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

$$C_\Delta = \frac{3}{8} = 0.375$$

summary
THREE CENTRAL QUANTITIES IN NETWORK SCIENCE

Degree distribution: $P(k)$

Path length: $<d>$

Clustering coefficient:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$
**Undirected**

\[
A_{ij} = \begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
\end{pmatrix}
\]

\[
A_{ii} = 0 \quad A_{ij} = A_{ji}
\]

\[
L = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij} \quad <k> = \frac{2L}{N}
\]

**Directed**

\[
A_{ij} = \begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{pmatrix}
\]

\[
A_{ii} = 0 \quad A_{ij} \neq A_{ji}
\]

\[
L = \sum_{i,j=1}^{N} A_{ij} \quad <k> = \frac{L}{N}
\]

Actor network, protein-protein interactions

WWW, citation networks
Unweighted (undirected)

\[ A_{ij} = \begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
\end{pmatrix} \]

\[ A_{ii} = 0 \quad A_{ij} = A_{ji} \]

\[ L = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij} \quad <k> = \frac{2L}{N} \]

protein-protein interactions, www

Weighted (undirected)

\[ A_{ij} = \begin{pmatrix}
0 & 2 & 0.5 & 0 \\
2 & 0 & 1 & 4 \\
0.5 & 1 & 0 & 0 \\
0 & 4 & 0 & 0 \\
\end{pmatrix} \]

\[ A_{ii} = 0 \quad A_{ij} = A_{ji} \]

\[ L = \frac{1}{2} \sum_{i,j=1}^{N} \text{nonzero}(A_{ij}) \quad <k> = \frac{2L}{N} \]

Call Graph, metabolic networks
**Self-interactions**

\[
A_{ij} = \begin{pmatrix}
1 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 \\
0 & 1 & 0 & 1
\end{pmatrix}
\]

- \( A_{ii} \neq 0 \)
- \( A_{ij} = A_{ji} \)

\[
L = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii}
\]

**Multigraph (undirected)**

\[
A_{ij} = \begin{pmatrix}
0 & 2 & 1 & 0 \\
2 & 0 & 1 & 3 \\
1 & 1 & 0 & 0 \\
0 & 3 & 0 & 0
\end{pmatrix}
\]

- \( A_{ii} = 0 \)
- \( A_{ij} = A_{ji} \)

\[
L = \frac{1}{2} \sum_{i,j=1}^{N} \text{nonzero}(A_{ij}) < k > = \frac{2L}{N}
\]

*Protein interaction network, www*

*Social networks, collaboration networks*
Complete Graph
(undirected)

\[
A_{ij} = \begin{pmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0
\end{pmatrix}
\]

\[A_{ii} = 0 \quad A_{ij} = 1\]

\[L = L_{\text{max}} = \frac{N(N-1)}{2} \quad <k> = N - 1\]

Actor network, protein-protein interactions
GRAPHOLOGY: Real networks can have multiple characteristics

WWW > directed multigraph with self-interactions

Protein Interactions > undirected unweighted with self-interactions

Collaboration network > undirected multigraph or weighted.

Mobile phone calls > directed, weighted.

Facebook Friendship links > undirected, unweighted.
THREE CENTRAL QUANTITIES IN NETWORK SCIENCE

A. Degree distribution:

\[ p_k = \{0.25; 0.5; 0.25\} \]

B. Path length:

\[ <d> = 1.33 \]

\[ d_{\text{max}} = 2 \]

\[ d_{1,2} = d_{2,3} = d_{2,4} = d_{3,4} = 1 \]
\[ d_{1,3} = d_{1,4} = 2 = d_{\text{max}} \]
\[ <d> = (4*1 + 2*2) / 6 = 1.33 \]

C. Clustering coefficient:

\[ C_1 = 0; \quad C_2 = 1/3 = 0.33 \]
\[ C_3 = C_4 = 1/1 = 1 \]
\[ <C> = 1.33 / 4 = 0.33 \]

\[ p_k = \{0.25; 0.5; 0.25\} \]
\[ <d> = 1.33 \]

\[ C_i = \frac{2e_i}{k_i(k_i - 1)} \]
protein-gene interactions

protein-protein interactions

GENOME

PROTEOME

METABOLISM

Bio-chemical reactions

Citrate Cycle
Metabolic Network
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

Undirected network
N=2,018 proteins as nodes
L=2,930 binding interactions as links.
Average degree $<k>=2.90$.

Not connected: 185 components
the largest (giant component) 1,647 nodes
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

Network Science: Graph Theory

\( p_k \) is the probability that a node has degree \( k \).

\[ N_k = \# \text{ nodes with degree } k \]

\[ N_k = N^* \, p_k \]

There is the same number of hubs with the increasing node degrees.
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

\[
d_{\text{max}} = 14
\]

\[
\langle d \rangle = 5.61
\]
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

\[ C_i = \frac{2e_i}{k_i(k_i - 1)} \]

\[ \langle C \rangle = 0.12 \]

Network Science: Graph Theory
### Node degrees, connected components (cc), and paths

<table>
<thead>
<tr>
<th>Node</th>
<th>deg</th>
<th>cc</th>
<th>paths</th>
<th>deg</th>
<th>cc</th>
<th>paths</th>
<th>deg</th>
<th>cc</th>
<th>paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>$3<em>1+2</em>2+2*3=13$</td>
<td>3</td>
<td>0.33</td>
<td>$3<em>1+2</em>2+2*3=13$</td>
<td>2</td>
<td>0</td>
<td>$2<em>1+2</em>2+3+4+5=18$</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>$3<em>1+2</em>2+2*3=13$</td>
<td>2</td>
<td>1</td>
<td>$2<em>1+2</em>2+2*3+4=16$</td>
<td>2</td>
<td>0</td>
<td>$2<em>1+2</em>2+3+4+5=22$</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.6</td>
<td>$5<em>1+2</em>2=9$</td>
<td>2</td>
<td>0</td>
<td>$2<em>1+4</em>2+3=13$</td>
<td>2</td>
<td>0</td>
<td>$2<em>1+2</em>2+3+4+5=16$</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.6</td>
<td>$5<em>1+2</em>2=9$</td>
<td>3</td>
<td>0.33</td>
<td>$3<em>1+2</em>2+2*3=13$</td>
<td>1</td>
<td>0</td>
<td>$1+2+3+4+5+6+7=28$</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.6</td>
<td>$5<em>1+2</em>2=9$</td>
<td>2</td>
<td>0</td>
<td>$2<em>1+4</em>2+3=13$</td>
<td>2</td>
<td>0</td>
<td>$2<em>1+2</em>2+3+4+5=16$</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0.6</td>
<td>$5<em>1+2</em>2=9$</td>
<td>3</td>
<td>0.33</td>
<td>$3<em>1+2</em>2+2*3=13$</td>
<td>2</td>
<td>0</td>
<td>$2<em>1+2</em>2+3+4+5=18$</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>$3<em>1+2</em>2+2*3=13$</td>
<td>3</td>
<td>0.33</td>
<td>$3<em>1+2</em>2+2*3=13$</td>
<td>1</td>
<td>0</td>
<td>$1+2+3+4+5+6+7=28$</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1</td>
<td>$3<em>1+2</em>2+2*3=13$</td>
<td>2</td>
<td>1</td>
<td>$2<em>1+2</em>2+2*3+4=16$</td>
<td>2</td>
<td>0</td>
<td>$2<em>1+2</em>3+4+5+6+7=22$</td>
</tr>
</tbody>
</table>

### Averages and Maximums

- **Average deg:** 4, **Average cc:** 0.8
- **Minimum paths:** $1.571428571$, **Maximum paths:** $1.928571429$
- **Minimum cc:** 0, **Minimum paths:** 3
- **Maximum deg:** 5, **Maximum cc:** 1, **Maximum paths:** 4

---

**Network Properties:**
- **N=8, E=16**
- **N=8, E=10**
- **N=8, E=7**

- All edges are bridges
- No cycles
Introduction
RANDOM NETWORK MODEL

Network Science: Random Networks
The random network model
Erdös-Rényi model (1960)

Connect with probability $p$

$p = \frac{1}{6}$ \hspace{0.5cm} N = 10

$\langle k \rangle \sim 1.5$
Definition:

A random graph is a graph of N nodes where each pair of nodes is connected by probability $p$. 

**$G(N, L)$ Model**

N labeled nodes are connected with $L$ randomly placed links. Erdős and Rényi used this definition in their string of papers on random networks [2-9].

**$G(N, p)$ Model**

Each pair of N labeled nodes is connected with probability $p$, a model introduced by Gilbert [10].
RANDOM NETWORK MODEL

$p=1/6$
$N=12$

$L=8$
$L=10$
$L=7$
p=0.03
N=100
The number of links is variable
RANDOM NETWORK MODEL

$p=1/6$
$N=12$

$L=8$
$L=10$
$L=7$
**P(L):** the probability to have exactly *L* links in a network of *N* nodes and probability *p*:

\[
P(L) = \binom{N}{2} p^L (1 - p)^{\frac{N(N-1)}{2} - L}
\]

- The maximum number of links in a network of *N* nodes.
- Number of different ways we can choose *L* links among all potential links.

Binomial distribution...
\[ P(x) = \binom{N}{x} p^x (1 - p)^{N-x} \]

\[ \langle x \rangle = Np \]

\[ \langle x^2 \rangle = p(1 - p)N + p^2 N^2 \]

\[ \sigma_x = \left( \langle k^2 \rangle - \langle k \rangle^2 \right)^{1/2} = \left[ p(1 - p)N \right]^{1/2} \]

**RANDOM NETWORK MODEL**

**\( P(L) \):** the probability to have a network of exactly \( L \) links

\[
P(L) = \binom{N}{L} p^L (1 - p)^{\frac{N(N-1)-L}{2}}
\]

• The average number of links \( \langle L \rangle \) in a random graph

\[
\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} LP(L) = p \frac{N(N-1)}{2}
\]

\[
\langle k \rangle = \frac{2L}{N} = p(N - 1)
\]

• The standard deviation

\[
\sigma^2 = p(1 - p) \frac{N(N-1)}{2}
\]
Degree distribution
As the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of \(<k>\).

\[ P(k) = \binom{N-1}{k} p^k (1 - p)^{(N-1)-k} \]

\[ \langle k \rangle = p(N-1) \]

\[ \sigma_k^2 = p(1 - p)(N-1) \]

\[ \frac{\sigma_k}{\langle k \rangle} = \left[ \frac{1 - p}{p} \frac{1}{(N-1)} \right]^{1/2} \approx \frac{1}{(N-1)^{1/2}} \]
DEGREE DISTRIBUTION OF A RANDOM GRAPH

\[ P(k) = \binom{N-1}{k} p^k (1 - p)^{(N-1) - k} \]

For large \( N \) and small \( k \), we can use the following approximations:

\[
\binom{N-1}{k} = \frac{(N-1)!}{k!(N-1-k)!} = \frac{(N-1)(N-1-1)(N-1-2) \ldots (N-1-k+1)(N-1-k)!}{k!(N-1-k)!} = \frac{(N-1)^k}{k!}
\]

\[
\ln[(1-p)^{(N-1) - k}] = (N-1-k)\ln(1 - \frac{<k>}{N-1}) = -(N-1-k) \frac{<k>}{N-1} = -<k>(1 - \frac{k}{N-1}) \approx -<k>
\]

\[
(1-p)^{(N-1) - k} = e^{-<k>}
\]

\[
\ln(1 + x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots \text{ for } |x| \leq 1
\]

\[
P(k) = \binom{N-1}{k} p^k (1 - p)^{(N-1) - k} = \frac{(N-1)^k}{k!} p^k e^{-<k>} = \frac{(N-1)^k}{k!} \left( \frac{<k>}{N-1} \right)^k e^{-<k>} = e^{-<k>} \frac{<k>^k}{k!}
\]
POISSON DEGREE DISTRIBUTION

For large $N$ and small $k$, we arrive at the Poisson distribution:

\[
P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k} \quad \quad \quad <k> = p(N-1) \quad \quad \quad p = \frac{<k>}{(N-1)}
\]

\[
P(k) = e^{-<k>} \frac{<k>^k}{k!}
\]
$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$

- Poisson
- Binomial

$N = 10^2$, $N = 10^3$, $N = 10^4$, $N = 10^5$
In the random graph literature, it is often assumed that the connection probability $p(N)$ scales as $N^z$, where $z$ is a tunable parameter between $-\infty$ and 0. In this language Erdős and Rényi discovered that as we vary $z$, key properties of random graphs appear quite suddenly.

A graph has a given property $Q$ if the probability of having $Q$ approaches 1 as $N \to \infty$. That is, for a given $z$ either almost every graph has the property $Q$ or almost no graph has it. For example, for $z$ less than $-3/2$ almost all graphs contain only isolated nodes and pairs of nodes connected by a link. Once $z$ exceeds $-3/2$, most networks will contain paths connecting three or more nodes, see image below.

\[ p = \frac{\langle k \rangle}{N-1} \approx \frac{\langle k \rangle}{N} = N^z \] so $\langle k \rangle = N^{z+1}$ and $z = \ln(\langle k \rangle) - 1$.
Real networks are supercritical
Section 7

Network Science: Random Networks

Subcritical | Supercritical | Fully Connected

Internet | | ✗ |

Power Grid | ✗ | |

Science Collaboration | | ✗ |

Actor Network | ✗ | |

Yeast Protein Interactions | | ✗ |

<table>
<thead>
<tr>
<th>Network</th>
<th>$N$</th>
<th>$L$</th>
<th>$&lt;k&gt;$</th>
<th>$\ln N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>192,244</td>
<td>609,066</td>
<td>6.34</td>
<td>12.17</td>
</tr>
<tr>
<td>Power Grid</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>8.51</td>
</tr>
<tr>
<td>Science Collaboration</td>
<td>23,133</td>
<td>186,936</td>
<td>8.08</td>
<td>10.04</td>
</tr>
<tr>
<td>Actor Network</td>
<td>212,250</td>
<td>3,054,278</td>
<td>28.78</td>
<td>12.27</td>
</tr>
<tr>
<td>Yeast Protein Interactions</td>
<td>2,018</td>
<td>2,930</td>
<td>2.90</td>
<td>7.61</td>
</tr>
</tbody>
</table>
Section 7

CONCLUSIONS

The measurements indicate that real networks extravagantly exceed the $k = 1$ threshold. Sociologists estimate that an average person has around 1,000 acquaintances; a typical neuron is connected to dozens of other neurons, some to thousands; in our cells, each molecule takes part in several chemical reactions, some, like water, in hundreds.

The average degree of real networks is well beyond the $k = 1$ threshold, implying they all have a giant component.

Do we have single component (if $k > \ln N$), or multiple components (if $k < \ln N$)? For social networks this requires $k \geq \ln(7 \times 10^9) \approx 22.7$; so nearly two dozens acquaintances per person; with $k \approx 1,000$ this is clearly satisfied. Most real networks do not satisfy this criteria, e.g., the Internet implying some routers are disconnected, so of little utility!

Most real networks are in the supercritical regime. This means that these networks have a giant component, but it coexists with many disconnected components and nodes, but only if real networks are accurately described by the Erdős-Rényi model, i.e. are random.

Today, we will further discuss the structure of real networks, we will understand why real networks can stay connected despite failing the $k > \ln N$ criteria.