Frontiers of Network Science
Fall 2020

Class 4: Graph Theory II
(Chapter 2 in Textbook)

Boleslaw Szymanski

based on slides by Albert-László Barabási & Roberta Sinatra
<table>
<thead>
<tr>
<th>NETWORK</th>
<th>NODES</th>
<th>LINKS</th>
<th>DIRECTED</th>
<th>UNDIRECTED</th>
<th>N</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>Routers</td>
<td>Internet connections</td>
<td>Undirected</td>
<td></td>
<td>192,244</td>
<td>609,066</td>
</tr>
<tr>
<td>WWW</td>
<td>Webpages</td>
<td>Links</td>
<td>Directed</td>
<td></td>
<td>325,729</td>
<td>1,497,134</td>
</tr>
<tr>
<td>Power Grid</td>
<td>Power plants, transformers</td>
<td>Cables</td>
<td>Undirected</td>
<td></td>
<td>4,941</td>
<td>6,594</td>
</tr>
<tr>
<td>Mobile Phone Calls</td>
<td>Subscribers</td>
<td>Calls</td>
<td>Directed</td>
<td></td>
<td>36,595</td>
<td>91,826</td>
</tr>
<tr>
<td>Email</td>
<td>Email addresses</td>
<td>Emails</td>
<td>Directed</td>
<td></td>
<td>57,194</td>
<td>103,731</td>
</tr>
<tr>
<td>Science Collaboration</td>
<td>Scientists</td>
<td>Co-authorship</td>
<td>Undirected</td>
<td></td>
<td>23,133</td>
<td>93,439</td>
</tr>
<tr>
<td>Actor Network</td>
<td>Actors</td>
<td>Co-acting</td>
<td>Undirected</td>
<td></td>
<td>702,388</td>
<td>29,397,908</td>
</tr>
<tr>
<td>Citation Network</td>
<td>Paper</td>
<td>Citations</td>
<td>Directed</td>
<td></td>
<td>449,673</td>
<td>4,689,479</td>
</tr>
<tr>
<td>E. Coli Metabolism</td>
<td>Metabolites</td>
<td>Chemical reactions</td>
<td>Directed</td>
<td></td>
<td>1,039</td>
<td>5,802</td>
</tr>
<tr>
<td>Protein Interactions</td>
<td>Proteins</td>
<td>Binding interactions</td>
<td>Undirected</td>
<td></td>
<td>2,018</td>
<td>2,930</td>
</tr>
</tbody>
</table>
Degree, Average Degree and Degree Distribution
A BIT OF STATISTICS

**BRIEF STATISTICS REVIEW**

Four key quantities characterize a sample of $N$ values $x_1, \ldots, x_N$:

**Average (mean):**

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

**The $n^{th}$ moment:**

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i^n$$

**Standard deviation:**

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \langle x \rangle)^2}$$

**Distribution of $x$:**

$$p_x = \frac{1}{N} \sum_{i} \delta_{x,x_i}$$

where $p_x$ follows

$$\sum_i p_x = 1 \left( \int p_x \, dx = 1 \right)$$
AVERAGE DEGREE

Undirected

\[ \langle k \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} k_i \]

\[ \langle k \rangle \equiv \frac{2L}{N} \]

N – the number of nodes in the graph

Directed

\[ \langle k^{in} \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} k_i^{in} \]
\[ \langle k^{out} \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} k_i^{out} \]
\[ \langle k^{in} \rangle = \langle k^{out} \rangle \]

\[ \langle k \rangle \equiv \frac{L}{N} \]
Degree distribution

$P(k)$: probability that a randomly chosen node has degree $k$

$N_k = \# \text{ nodes with degree } k$

$P(k) = \frac{N_k}{N}$  plot
DEGREE DISTRIBUTION

(a) Degree distribution plot showing the distribution of degrees ($k$) and the corresponding probability ($P_k$).

(b) Network visualization with highlighted hubs.

(c) Log-log plot of degree distribution ($P_k$ vs. $k$) indicating a power-law distribution with hubs.
**Discrete Representation:** $p_k$ is the probability that a node has degree $k$.

**Continuum Description:** $p(k)$ is the pdf of the degrees, where

$$\int_{k_1}^{k_2} p(k) \, dk$$

represents the probability that a node’s degree is between $k_1$ and $k_2$.

**Normalization condition:**

$$\sum_{0}^{\infty} p_k = 1$$

$$\int_{K_{\min}}^{\infty} p(k) \, dk = 1$$

where $K_{\min}$ is the minimal degree in the network.
Adjacency matrix
\[ A_{ij} = \begin{cases} 1 & \text{if there is a link between node } i \text{ and } j \\ 0 & \text{if nodes } i \text{ and } j \text{ are not connected to each other.} \end{cases} \]

Note that for a directed graph (right) the matrix is not symmetric.

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

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

\[ A_{ij} = 1 \text{ if there is a link pointing from node } j \text{ and } i \]
\[ A_{ij} = 0 \text{ if there is no link pointing from } j \text{ to } i. \]
**Adjacency Matrix and Node Degrees**

**Undirected**

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

\[ k_i = \sum_{j=1}^{N} A_{ij} \]

\[ k_j = \sum_{i=1}^{N} A_{ij} \]

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

**Directed**

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

\[ k_i^{\text{in}} = \sum_{j=1}^{N} A_{ij} \]

\[ k_j^{\text{out}} = \sum_{i=1}^{N} A_{ij} \]

\[ L = \sum_{i=1}^{N} k_i^{\text{in}} = \sum_{j=1}^{N} k_j^{\text{out}} = \sum_{i,j} A_{ij} \]

\[ A_{ij} \neq A_{ji} \]

\[ A_{ii} = 0 \]
### Adjacency Matrix

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>h</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Real networks are sparse
The maximum number of links a network of N nodes can have is:

\[ L_{\text{max}} = \binom{N}{2} = \frac{N(N - 1)}{2} \]

A graph with degree \( L = L_{\text{max}} \) is called a complete graph, and its average degree is \( \langle k \rangle = N - 1 \)
Most networks observed in real systems are sparse:

$$L \ll L_{\text{max}}$$

or

$$\langle k \rangle \ll N-1.$$ 

WWW (ND Sample): $N=325,729; \ L=1.4 \ 10^6 \ \ L_{\text{max}}=10^{12} \ \ <k>=4.51$

Protein (S. Cerevisiae): $N=1,870; \ L=4,470 \ \ L_{\text{max}}=10^7 \ \ <k>=2.39$

Coauthorship (Math): $N=70,975; \ L=2 \ 10^5 \ \ L_{\text{max}}=3 \ 10^{10} \ \ <k>=3.9$

Movie Actors: $N=212,250; \ L=6 \ 10^6 \ \ L_{\text{max}}=1.8 \ 10^{13} \ \ <k>=28.78$

(Source: Albert, Barabasi, RMP2002)
ADJACENCY MATRICES ARE SPARSE
WEIGHTED AND UNWEIGHTED NETWORKS
$A_{ij} = w_{ij}$
**Unweighted (undirected)**

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

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

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

**Weighted (undirected)**

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

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

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

protein-protein interactions, www

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)**

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

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

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

\[ <k> = \frac{2L}{N} \]

*Network Science: Graph Theory*

Protein interaction network, www

Social networks, collaboration networks
Complete Graph
(undirected)

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

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

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

Actor network, protein-protein interactions
The maximum number of links a network of $N$ nodes can have is:

$$L_{\text{max}} = \binom{N}{2} = \frac{N(N-1)}{2}$$
Section 2.7

BIPARTITE NETWORKS
A bipartite graph (or bigraph) is a graph whose nodes can be divided into two disjoint sets $U$ and $V$ such that every link connects a node in $U$ to one in $V$; that is, $U$ and $V$ are independent sets.

Examples:

- Hollywood actor network
- Collaboration networks
- Disease network (diseasome)
Gene network

Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)
PATHOLOGY
A path is a sequence of nodes in which each node is adjacent to the next one.  

\[ P_{i_0,i_n} \text{ of length } n \text{ between nodes } i_0 \text{ and } i_n \text{ is an ordered collection of } n+1 \text{ nodes and } n \text{ links} \]

\[ P_n = \{i_0, i_1, i_2, \ldots, i_n\} \quad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\} \]

- In a directed network, the path can follow only the direction of an arrow.
The distance (shortest path, geodesic path) between two nodes is defined as the number of edges along the shortest path connecting them.

*If the two nodes are disconnected, the distance is infinity.

In directed graphs each path needs to follow the direction of the arrows. Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).
N_{ij}, number of paths between any two nodes i and j:

**Length n=1:** If there is a link between i and j, then A_{ij} = 1 and A_{ij} = 0 otherwise.

**Length n=2:** If there is a path of length two between i and j, then A_{ik}A_{kj} = 1, and A_{ik}A_{kj} = 0 otherwise.

The number of paths of length 2:

\[ N^{(2)}_{ij} = \sum_{k=1}^{N} A_{ik}A_{kj} = [A^2]_{ij} \]

**Length n:** In general, if there is a path of length n between i and j, then A_{ik}…A_{ij} = 1 and A_{ik}…A_{ij} = 0 otherwise.

The number of paths of length n between i and j is *

\[ N^{(n)}_{ij} = [A^n]_{ij} \]

* holds for both directed and undirected networks.
Distance between node 0 and node 4:

1. Start at 0.
Distance between node 0 and node 4:
1. Start at 0.
2. Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.
Distance between node 0 and node 4:
1. Start at 0.
2. Find the nodes adjacent to 0. Mark them as at distance 1. Put them in a queue.
3. Take the first node out of the queue. Find the unmarked nodes adjacent to it in the graph. Mark them with the label of 2. Put them in the queue.
Distance between node 0 and node 4:

1. Repeat until you find node 4 or there are no more nodes in the queue.
2. The distance between 0 and 4 is the label of 4 or, if 4 does not have a label, infinity.
Diameter: $d_{\text{max}}$ the maximum distance between any pair of nodes in the graph.

Average path length/distance, $\langle d \rangle$, for a connected graph:

$$\langle d \rangle \equiv \frac{1}{\sum_{i,j \neq i} 2L_{\text{max}}} \sum_{i,j \neq i} d_{ij}$$

where $d_{ij}$ is the distance from node $i$ to node $j$

In an undirected graph $d_{ij} = d_{ji}$, so we only need to count them once:

$$\langle d \rangle \equiv \frac{1}{L_{\text{max}}} \sum_{i,j > i} d_{ij}$$
The path with the shortest length between two nodes (distance).
**Diameter**

The longest shortest path in a graph

\[ l_{1 \rightarrow 4} = 3 \]

**Average Path Length**

The average of the shortest paths for all pairs of nodes.

\[ \frac{l_{1 \rightarrow 2} + l_{1 \rightarrow 3} + l_{1 \rightarrow 4} + l_{1 \rightarrow 5} + l_{2 \rightarrow 3} + l_{2 \rightarrow 4} + l_{2 \rightarrow 5} + l_{3 \rightarrow 4} + l_{3 \rightarrow 5} + l_{4 \rightarrow 5}}{10} = 1.6 \]
PATHOLOGY: summary

**Cycle**
A path with the same start and end node.

**Self-avoiding Path**
A path that does not intersect itself.
Eulerian Path

A path that traverses each link exactly once.

Hamiltonian Path

A path that visits each node exactly once.