Basic Network Properties and the Random Graph Model
Review of basic probability:
- Today, Thu 9/29
- In Gates B01, 4-6pm

Review of basic linear algebra:
- Tomorrow, Fri 9/30
- Gates B03, 4-6pm

Next week:
- Intro to SNAP (Gates B01, 4-6pm on Thu 10/6)
- Intro to NetworkX (Gates B03, 4-6pm on Fri 10/7)
Recall from the last lecture:

1) We took a real system: the Web
2) We represented it as a directed graph
3) We used the language of graph theory
   - Strongly Connected Components
4) We designed a computational experiment:
   - Find In- and Out-components of a given node \( v \)
5) We learned something about the structure of the Web

This class:

- Define basic terminology and measures that you can compute on networks
Undirected vs. Directed Networks

**Undirected**
- Links: undirected (symmetrical)
  - Undirected links:
    - Collaborations
    - Friendship on Facebook

**Directed**
- Links: directed (arcs)
  - Directed links:
    - Phone calls
    - Following on Twitter
Adjacency Matrix

$A_{ij} = 1$ if there is a link between node $i$ and $j$

$A_{ij} = 0$ if nodes $i$ and $j$ are not connected to each other

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

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

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

Undirected

Node degree: the number of links connected to the node

\[ k_i = 4 \]

Avg. degree: \[ \bar{k} = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2E}{N} \]

Directed

In directed networks we define an in-degree and out-degree. The (total) degree of a node is the sum of in- and out-degree.

\[ k^\text{in}_C = 2 \quad k^\text{out}_C = 1 \quad k_C = 3 \]

Source: A node with \( k^\text{in} = 0 \)
Sink: A node with \( k^\text{out} = 0 \)
The maximum number of edges in an undirected graph on $N$ nodes is

$$E_{\text{max}} = \binom{N}{2} = \frac{N(N-1)}{2}$$

A graph with the number of edges $E = E_{\text{max}}$ is a complete graph, and its average degree is $N-1$. 
Most real-world networks are sparse

\[ E \ll E_{\text{max}} \quad \text{or} \quad \bar{k} \ll N-1 \]

<table>
<thead>
<tr>
<th>Network Type</th>
<th>N</th>
<th>\langle k \rangle</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW (Stanford-Berkeley)</td>
<td>319,717</td>
<td>9.65</td>
</tr>
<tr>
<td>Social networks (LinkedIn)</td>
<td>6,946,668</td>
<td>8.87</td>
</tr>
<tr>
<td>Communication (MSN IM)</td>
<td>242,720,596</td>
<td>11.1</td>
</tr>
<tr>
<td>Coauthorships (DBLP)</td>
<td>317,080</td>
<td>6.62</td>
</tr>
<tr>
<td>Internet (AS-Skitter)</td>
<td>1,719,037</td>
<td>14.91</td>
</tr>
<tr>
<td>Roads (California)</td>
<td>1,957,027</td>
<td>2.82</td>
</tr>
<tr>
<td>Protein (S. Cerevisia)</td>
<td>1,870</td>
<td>2.39</td>
</tr>
</tbody>
</table>

(Source: Leskovec et al., Internet Mathematics, 2009)

Consequence: Adjacency matrix is filled with zeros!

(Density \(E/N^2\): WWW = 1.51 \times 10^{-5}, MSN IM = 2.27 \times 10^{-8})
More Types of Graphs:

- **Unweighted** (undirected)

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

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

  \[ E = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij} \quad \bar{k} = \frac{2E}{N} \]

  Friendships, WWW

- **Weighted** (undirected)

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

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

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

  Call graph, Email graph
More Types of Graphs:

- **Self-edges**
  (undirected)

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

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

- **Multigraph**
  (undirected)

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

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

  \[k = \frac{2E}{N}\]

WWW, Email

Social networks, collaboration networks
Network Representations

WWW >> directed multigraph with self-interactions

Facebook friendships >> undirected, unweighted

Citation networks >> unweighted directed acyclic

Collaboration networks >> undirected multigraph or weighted

Mobile phone calls >> directed, (weighted?) multigraph

Protein Interactions >> undirected, unweighted with self-interactions
Bipartite graph 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:
- Authors-to-papers
- Movies-to-Actors
- Users-to-Movies

“Folded” networks
- Author collaboration networks
- Actor collaboration networks
Network Properties: How to Characterize a Network?
Degree Distribution

- **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
A path is a sequence of nodes in which each node is adjacent to the next one

\[ P_n = \{i_0, i_1, i_2, \ldots, i_n\} \]

Path can intersect itself and pass through the same edge multiple times

- E.g.: ACBDCDEG
- In a directed graph a path can only follow the direction of the “arrow”
Number of Paths

- **Number of paths between nodes \( u \) and \( v \):**
  - **Length \( h=1 \):** If there is a link between \( u \) and \( v \),
    \[ A_{uv} = 1 \text{ else } A_{uv} = 0 \]
  - **Length \( h=2 \):** If there is a path of length two
    between \( u \) and \( v \) then \( A_{uk} A_{kv} = 1 \text{ else } A_{uk} A_{kv} = 0 \)
    \[ H^{(2)}_{uv} = \sum_{k=1}^{N} A_{uk} A_{kv} = [A^2]_{uv} \]
  - **Length \( h \):** If there is a path of length \( h \) between \( u \) and \( v \) then
    \( A_{uk} \ldots A_{kv} = 1 \text{ else } A_{uk} \ldots A_{kv} = 0 \)
    So, the no. of paths of length \( h \) between \( u \) and \( v \) is
    \[ H^{(h)}_{uv} = [A^h]_{uv} \]
    (holds for both directed and undirected graphs)
Distance (shortest path, geodesic) between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes.

- If the two nodes are disconnected, the distance is defined as infinite.

In directed graphs paths need to follow the direction of the arrows.

- Consequence: Distance is not symmetric: $h(A, C) \neq h(C, A)$
Finding Shortest Paths

- **Breath-First Search:**
  - Start with node $u$, mark it to be at distance $h_u(u)=0$, add $u$ to the queue
  - While queue not empty:
    - Take node $v$ off the queue, put its unmarked neighbor $w$ into the queue and mark $h_u(w)=h_u(v)+1$
Network Diameter

- **Diameter**: the maximum (shortest path) distance between any pair of nodes in the graph
- **Average path length/distance** for a connected graph (component) or a strongly connected (component of a) directed graph

\[
\bar{h} = \frac{1}{2E_{\text{max}}} \sum_{i,j \neq i} h_{ij}
\]

where \( h_{ij} \) is the distance from node \( i \) to node \( j \)

- Many times we compute the average only over the connected pairs of nodes (i.e., we ignore “infinite” paths)
Clustering Coefficient

- **Clustering coefficient:**
  - What portion of $i$’s neighbors are connected?
  - Node $i$ with degree $k_i$
  - $C_i \in [0,1]$

- $C_i = \frac{2e_i}{k_i(k_i-1)}$ where $e_i$ is the number of edges between the neighbors of node $i$

- **Average Clustering Coefficient:** $C = \frac{1}{N} \sum_{i}^{N} C_i$
Clustering Coefficient

Clustering coefficient:

- What portion of $i$’s neighbors are connected?
- Node $i$ with degree $k_i$

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

where $e_i$ is the number of edges between the neighbors of node $i$

$k_B = 2, \quad e_B = 1, \quad C_B = 2/2 = 1$

$k_D = 4, \quad e_D = 2, \quad C_D = 4/12 = 1/3$
Key Network Properties

Degree distribution: $P(k)$

Path length: $h$

Clustering coefficient: $C$
**Regular Lattice: 1D**

- \( P(k) = \delta(k-4) \quad k=4 \) for each node
- \( C = \frac{1}{2} \) for each node if \( N > 6 \)
- Path length:

  \[ h_{\text{max}} \approx \frac{N}{2} \]

  Alternative calculation:

  \[ \sum_{h=1}^{h_{\text{max}}} 4 \approx N \quad \Rightarrow \quad h_{\text{max}} \approx \frac{N}{4} \]

- The average path-length is \( \bar{h} \approx N \)
- Constant degree, constant clustering coefficient.
Regular Lattice: 2D

- $P(k) = \delta(k-6)$
  - $k=6$ for each inside node
- $C = 6/15$ for inside nodes
- Path length:
  
  $\sum_{h=1}^{h_{\text{max}}} 6h \approx N \Rightarrow h_{\text{max}} \propto \sqrt{N}$

- In general, for lattices:
  - average path-length is $\bar{h} \approx N^{1/D}$
  - Constant degree, constant clustering coefficient
### 3-way Cayley Tree

- **Degree:** \( \bar{k} = 2 \)
  - \( k=3 \) for non-leaves
  - \( k=1 \) for leaves
- **\( C = 0 \)**
- **Path length:**

\[
3 \sum_{h=1}^{h_{\text{max}}} 2^{h-1} \approx N \quad \Rightarrow \quad h_{\text{max}} \propto \log_{\bar{k}} N = \frac{\log N}{\log \bar{k}}
\]

- Distances vary logarithmically with \( N \).

Constant degree, no clustering.
Erdös-Rényi Random Graph Model
Erdös-Renyi Random Graph [Erdös-Renyi, ‘60]

Two variants:

- $G_{n,p}$: undirected graph on $n$ nodes and each edge $(u,v)$ appears i.i.d. with probability $p$
- $G_{n,m}$: undirected graph with $n$ nodes, and $m$ uniformly at random picked edges

What kinds of networks does such model produce?
Random Graph Model

- $n$ and $p$ do not uniquely define the graph
- We can have many different realizations. How many?

The probability of $G_{np}$ to form a particular graph $G(N,E)$ is

$$P(G(N,E)) = p^E (1-p)^{N(N-1)/2-E}$$

That is, each concrete graph $G(N,E)$ appears with probability $P(G(N,E))$. 

$n = 10$  
p = 1/6
How many likely is a graph on $E$ edges?

$P(E)$: the probability that a given $G_{np}$ generates a graph on exactly $E$ edges:

$$P(E) = \binom{E_{\text{max}}}{E} p^E (1-p)^{E_{\text{max}}-E}$$

where $E_{\text{max}} = n(n-1)/2$ is the maximum possible number of edges

Binomial distribution >>>>
What is expected degree of a node?

Let $X_v$ be a random var. measuring the degree of the node $v$: 
$$E[X_v] = \sum_{j=0}^{n-1} j P(X_v = j)$$

- **Linearity of expectation:**
  - For any random variables $Y_1, Y_2, \ldots, Y_k$
  - If $Y=Y_1+Y_2+\ldots+Y_k$, then $E[Y] = \sum_i E[Y_i]$

Easier way:

- Decompose $X_v$ in $X_v = X_{v1} + X_{v2} + \ldots + X_{vn-1}$
  - where $X_{vu}$ is a $\{0,1\}$-random variable which tells if edge $(v,u)$ exists or not

$$E[X_v] = \sum_{y=1}^{n-1} E[X_{vu}] = (n-1)p$$

How to think about this?
- Prob. of node $u$ linking to node $v$ is $p$
- $u$ can link (flips a coin) for all other $(n-1)$ nodes
- Thus, the expected degree of node $u$ is $p(n-1)$
Degree Distribution

- Degree distribution of $G_{np}$ is Binomial.
- Let $P(k)$ denote a fraction of nodes with degree $k$:

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Select $k$ nodes from $n-1$

Probability of having $k$ edges

Probability of missing $n-1-k$ edges

$$\bar{k} = p(n-1)$$

$$\sigma_k^2 = p(1-p)(n-1)$$

$$\frac{\sigma_k}{\bar{k}} = \left[ \frac{1-p}{p} \frac{1}{(n-1)} \right]^{1/2} \approx \frac{1}{(n-1)^{1/2}}$$

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 $\bar{k}$. 

9/29/2011


35
Since edges in $G_{np}$ appear i.i.d with probability $p$

$$e_i \approx p \frac{k_i(k_i - 1)}{2} \quad \Rightarrow \quad C \approx p = \frac{\bar{k}}{N}$$

Clustering coefficient of a random graph is small. For a fixed degree $C$ decreases with the graph size $N$. 

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$
**Configuration model:**

- Assume a degree sequence $k_1, k_2, \ldots, k_N$
- Useful for as a “null” model of networks
  - We can compare the real network $G$ and a “random” graph $G'$ which has the same degree sequence as $G$