Measuring Networks and the Random Graph Model
How the Class Fits Together

**Measurements**
- Small diameter, Edge clustering
- Patterns of signed edge creation
- Viral Marketing, Blogosphere, Memetracking
- Scale-Free
- Densification power law, Shrinking diameters
- Strength of weak ties, Core-periphery

**Models**
- Erdös-Renyi model, Small-world model
- Structural balance, Theory of status
- Independent cascade model, Game theoretic model
- Preferential attachment, Copying model
- Microscopic model of evolving networks
- Kronecker Graphs

**Algorithms**
- Decentralized search
- Models for predicting edge signs
- Influence maximization, Outbreak detection, LIM
- PageRank, Hubs and authorities
- Link prediction, Supervised random walks
- Community detection: Girvan-Newman, Modularity
Choice of the proper network representation of a given system determines our ability to use networks successfully.
Directed vs. Undirected Graphs

Undirected

- **Links**: undirected (symmetrical, reciprocal)

  - Examples:
    - Collaborations
    - Friendship on Facebook

Directed

- **Links**: directed (arcs)

  - Examples:
    - Phone calls
    - Following on Twitter
Node Degrees

Node degree, $k_i$: the number of edges adjacent to node $i$
\[ k_A = 4 \]

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

In directed networks we define an **in-degree** and **out-degree**.
The (total) degree of a node is the sum of in- and out-degrees.
\[ k_{C}^{in} = 2 \quad k_{C}^{out} = 1 \quad k_{C} = 3 \]

**Source:** Node with $k^{in} = 0$
**Sink:** Node with $k^{out} = 0$
The **maximum number of edges** in an undirected graph on $N$ nodes is

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

An undirected graph with the number of edges $E = E_{\text{max}}$ is called a **complete graph**, and its average degree is $N-1$.
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 (they authored)
- Actors-to-Movies (they appeared in)
- Users-to-Movies (they rated)

“Folded” networks:
- Author collaboration networks
- Movie co-rating networks
Representing Graphs: Adjacency Matrix

\[ A_{ij} = 1 \] if there is a link from node \( i \) to node \( j \)
\[ A_{ij} = 0 \] otherwise

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

Note that for a directed graph (right) the matrix is not symmetric.
Represent graph as a set of edges:

- (2, 3)
- (2, 4)
- (3, 2)
- (3, 4)
- (4, 5)
- (5, 2)
- (5, 1)
**Adjacency list:**

- Easier to work with if network is **Large**
- Easier to work with if network is **Sparse**
- Allows us to quickly retrieve all neighbors of a given node
  - 1:
  - 2: 3, 4
  - 3: 2, 4
  - 4: 5
  - 5: 1, 2
Most real-world networks are **sparse**

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

- **WWW (Stanford-Berkeley):** \( N=319,717 \)  \( \langle k \rangle = 9.65 \)
- **Social networks (LinkedIn):** \( N=6,946,668 \)  \( \langle k \rangle = 8.87 \)
- **Communication (MSN IM):** \( N=242,720,596 \)  \( \langle k \rangle = 11.1 \)
- **Coauthorships (DBLP):** \( N=317,080 \)  \( \langle k \rangle = 6.62 \)
- **Internet (AS-Skitter):** \( N=1,719,037 \)  \( \langle k \rangle = 14.91 \)
- **Roads (California):** \( N=1,957,027 \)  \( \langle k \rangle = 2.82 \)
- **Proteins (S. Cerevisiae):** \( N=1,870 \)  \( \langle k \rangle = 2.39 \)

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

**Consequence:** Adjacency matrix is filled with zeros!

(Density of the matrix \( E/N^2 \): WWW = \( 1.51 \times 10^{-5} \), MSN IM = \( 2.27 \times 10^{-8} \))
Possible options:

- Weight (e.g. frequency of communication)
- Ranking (best friend, second best friend...)
- Type (friend, relative, co-worker)
- Sign: Friend vs. Foe, Trust vs. Distrust
- Properties depending on the structure of the rest of the graph: number of common friends
More Types of Graphs

- **Unweighted**
  (undirected)

- **Weighted**
  (undirected)

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

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

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

**Examples:** Friendship, Hyperlink

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

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

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

**Examples:** Collaboration, Internet, Roads
More Types of Graphs

- **Self-edges (self-loops)**
  (undirected)

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

  \(A_{ii} \neq 0\)

  \(A_{ij} = A_{ji}\)

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

  **Examples:** Proteins, Hyperlinks

- **Multigraph**
  (undirected)

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

  \(A_{ii} = 0\)

  \(A_{ij} = A_{ji}\)

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

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

  **Examples:** Communication, Collaboration
### Connectivity of Undirected Graphs

- **Connected (undirected) graph:**
  - Any two vertices can be joined by a path
  - A disconnected graph is made up by two or more connected components

![Graph](image)

**Bridge edge:** If we erase it, the graph becomes disconnected.

**Articulation point:** If we erase it, the graph becomes disconnected.
Connectivity of Directed Graphs

- **Strongly connected directed graph**
  - has a path from each node to every other node and vice versa (e.g., A-B path and B-A path)

- **Weakly connected directed graph**
  - is connected if we disregard the edge directions

Graph on the left is connected but not strongly connected (e.g., there is no way to get from F to G by following the edge directions).
Network Representations

WWW $\gg$ directed multigraph with self-edges

Facebook friendships $\gg$ undirected, unweighted

Citation networks $\gg$ unweighted, directed, acyclic

Collaboration networks $\gg$ undirected multigraph or weighted graph

Mobile phone calls $\gg$ directed, (weighted?) multigraph

Protein Interactions $\gg$ undirected, unweighted with self-interactions
Web as a Graph
Today we will talk about observations and models for the Web graph:

1) We will take a real system: the Web
2) We will represent it as a directed graph
3) We will use the language of graph theory
   - Strongly Connected Components
4) We will design a computational experiment:
   - Find In- and Out-components of a given node $v$
5) We will learn something about the structure of the Web: BOWTIE!
Q: What does the Web “look like” at a global level?

- **Web as a graph:**
  - Nodes = web pages
  - Edges = hyperlinks

- **Side issue:** What is a node?
  - Dynamic pages created on the fly
  - “dark matter” – inaccessible database generated pages
I teach a class on Networks.

CS224W: Classes are in the Gates building

Computer Science Department at Stanford

Stanford University
In early days of the Web links were navigational
Today many links are transactional
The Web as a Directed Graph

I'm a student at Univ. of X

My song lyrics

Classes

I teach at Univ. of X

Networks

Networks class blog

Blog post about Company Z

Blog post about college rankings

Uni. of X

I'm applying to college

USNews College Rankings

USNews Featured Colleges
Other Information Networks

Citations

References in an Encyclopedia
What Does the Web Look Like?

- How is the Web linked?
- What is the “map” of the Web?

Web as a directed graph [Broder et al. 2000]:

- Given node $v$, what can $v$ reach?
- What other nodes can reach $v$?

$In(v) = \{w \mid w \text{ can reach } v\}$

$Out(v) = \{w \mid v \text{ can reach } w\}$

For example:
$In(A) = \{A,B,C,E,G\}$
$Out(A) = \{A,B,C,D,F\}$
Two types of directed graphs:

- **Strongly connected:**
  - Any node can reach any node via a directed path
  
  \[ \text{In}(A) = \text{Out}(A) = \{A, B, C, D, E\} \]

- **Directed Acyclic Graph (DAG):**
  - Has no cycles: if \( u \) can reach \( v \), then \( v \) cannot reach \( u \)

Any directed graph can be expressed in terms of these two types!


A Strongly Connected Component (SCC) is a set of nodes $S$ so that:

- Every pair of nodes in $S$ can reach each other
- There is no larger set containing $S$ with this property

Strongly connected components of the graph: \{A,B,C,G\}, \{D\}, \{E\}, \{F\}
Fact: Every directed graph is a DAG on its SCCs

1. SCCs partitions the nodes of $G$
   - That is, each node is in exactly one SCC

2. If we build a graph $G'$ whose nodes are SCCs, and with an edge between nodes of $G'$ if there is an edge between corresponding SCCs in $G$, then $G'$ is a DAG

(1) Strongly connected components of graph $G$: \{A,B,C,G\}, \{D\}, \{E\}, \{F\}

(2) $G'$ is a DAG
Claim: SCCs partition nodes of $G$.

- This means: Each node is member of exactly 1 SCC

Proof by contradiction:

- Suppose there exists a node $v$ which is a member of two SCCs $S$ and $S'$

But then $S \cup S'$ is one large SCC!

Contradiction: By definition SCC is a maximal set with the SCC property, so $S$ and $S'$ are not two SCCs.
Claim: $G'$ (graph of SCCs) is a DAG.

- This means: $G'$ has no cycles

Proof by contradiction:

- Assume $G'$ is not a DAG
- Then $G'$ has a directed cycle
- Now all nodes on the cycle are mutually reachable, and all are part of the same SCC
- But then $G'$ is not a graph of connections between SCCs (SCCs are defined as maximal sets)
- Contradiction!

Now $\{A,B,C,G,E,F\}$ is a SCC!
Goal: Take a large snapshot of the Web and try to understand how its SCCs “fit together” as a DAG

Computational issue:

- Want to find a SCC containing node $v$?
- Observation:
  - $Out(v)$ ... nodes that can be reached from $v$
  - SCC containing $v$ is: $Out(v) \cap In(v)$
    
    $= Out(v, G) \cap Out(v, \overline{G})$, where $\overline{G}$ is $G$ with all edge directions flipped
**Example:**

- \( \text{Out}(A) = \{A, B, D, E, F, G, H\} \)
- \( \text{In}(A) = \{A, B, C, D, E\} \)
- So, \( \text{SCC}(A) = \text{Out}(A) \cap \text{In}(A) = \{A, B, D, E\} \)
There is a single giant SCC

- That is, there won’t be two SCCs

Heuristic argument:

- It just takes 1 page from one SCC to link to the other SCC
- If the 2 SCCs have millions of pages the likelihood of this not happening is very very small
**Structure of the Web**

- **Broder et al., 2000:**
  - Altavista crawl from October 1999
    - 203 million URLs
    - 1.5 billion links
  - Computer: Server with 12GB of memory

- **Undirected version of the Web graph:**
  - 91% nodes in the largest weakly conn. component
  - Are hubs making the web graph connected?
    - Even if they deleted links to pages with in-degree >10
      WCC was still ≈50% of the graph
Directed version of the Web graph:

- **Largest SCC:** 28% of the nodes (56 million)
- Taking a random node $v$
  - $\text{Out}(v) \approx 50\%$ (100 million)
  - $\text{In}(v) \approx 50\%$ (100 million)

What does this tell us about the conceptual picture of the Web graph?
The Bowtie Structure of the Web

203 million pages, 1.5 billion links [Broder et al. 2000]
What did we learn:
- Conceptual organization of the Web (i.e., the bowtie)

What did we not learn:
- Treats all pages as equal
  - Google’s homepage == my homepage
- What are the most important pages
  - How many pages have $k$ in-links as a function of $k$?
    The degree distribution: $\sim k^{-2}$
- Internal structure inside giant SCC
  - Clusters, implicit communities?
- How far apart are nodes in the giant SCC:
  - Distance = # of edges in shortest path
  - Avg. = 16 [Broder et al.]
Network Properties: How to Measure a Network?
Plan: Key Network Properties

Degree distribution: \( P(k) \)

Path length: \( h \)

Clustering coefficient: \( C \)
(1) Degree Distribution

- Degree distribution $P(k)$: Probability that a randomly chosen node has degree $k$
  
  $N_k = \# \text{ nodes with degree } k$

- Normalized histogram:
  
  $P(k) = \frac{N_k}{N} \rightarrow \text{ plot}$
(2) Paths in a Graph

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

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

- 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 $v$, $A_{uv} = 1$ else $A_{uv} = 0$
  - **Length $h=2$**: If there is a path of length two between $u$ and $v$ then $A_{uk}A_{kv} = 1$ else $A_{uk}A_{kv} = 0$
  - **Length $h$**: If there is a path of length $h$ between $u$ and $v$ then $A_{uk} \ldots A_{kv} = 1$ else $A_{uk} \ldots A_{kv} = 0$

So, the no. of paths of length $h$ between $u$ and $v$ is

$$H_{uv}^{(h)} = [A^h]_{uv}$$

(holds for both directed and undirected graphs)
Distance in a Graph

- **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 usually defined as infinite.*

- **In directed graphs** paths need to follow the direction of the arrows.
  - **Consequence:** Distance is **not symmetric**: $h_{A,C} \neq h_{C,A}$.
Network Diameter

- **Diameter**: the maximum (shortest path) distance between any pair of nodes in a graph

- **Average path length** 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 (that is, we ignore “infinite” length paths)
Breadth First Search:

- Start with node $u$, mark it to be at distance $h_u(u) = 0$, add $u$ to the queue
- While the queue not empty:
  - Take node $v$ off the queue, put its unmarked neighbors $w$ into the queue and mark $h_u(w) = h_u(v) + 1$
(3) 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 $i$

- $k_B = 2$, $e_B = 1$, $C_B = \frac{2}{2} = 1$
- $k_D = 4$, $e_D = 2$, $C_D = \frac{4}{12} = \frac{1}{3}$
Summary: Key Network Properties

Degree distribution: \[ P(k) \]

Path length: \[ h \]

Clustering coefficient: \[ C \]
Let’s measure $P(k)$, $h$ and $C$ on a real-world network!
MSN Messenger activity in June 2006:

- 245 million users logged in
- 180 million users engaged in conversations
- More than 30 billion conversations
- More than 255 billion exchanged messages
Communication: Geography
Network: 180M people, 1.3B edges
Messaging as a Multigraph

- Edge \((u,v)\) if users \(u\) and \(v\) exchanged at least 1 msg
- \(N=180\) million people
- \(E=1.3\) billion edges
MSN: (1) Connectivity

The graph shows the distribution of weakly connected component sizes. The x-axis represents the weakly connected component size, while the y-axis represents the count. The largest component contains 99.9% of the nodes.
MSN: (2) Degree Distribution

![Degree Distribution Graph]

- Count, $P(k)$
- Degree, $k$
- $3.5e+007$
- $3e+007$
- $2.5e+007$
- $2e+007$
- $1.5e+007$
- $1e+007$
- $5e+006$
- $0$
Note: We plotted the same data as on the previous slide, just the axes are now logarithmic.
MSN: (3) Clustering

Avg. clustering of the MSN:
$C = 0.1140$

$c$ (Clustering coefficient)

$k$ (Degree)

$C_k$: average $C_i$ of nodes $i$ of degree $k$:

$$C_k = \frac{1}{N_k \sum_{i:k_i=k} C_i}$$
MSN: (4) Diameter

Avg. path length 6.6
90% of the nodes can be reached in < 8 hops

Number of links between pairs of nodes

<table>
<thead>
<tr>
<th>Steps</th>
<th>#Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>3,96</td>
</tr>
<tr>
<td>4</td>
<td>8,648</td>
</tr>
<tr>
<td>5</td>
<td>3,299,252</td>
</tr>
<tr>
<td>6</td>
<td>28,395,849</td>
</tr>
<tr>
<td>7</td>
<td>79,059,497</td>
</tr>
<tr>
<td>8</td>
<td>52,995,778</td>
</tr>
<tr>
<td>9</td>
<td>10,321,008</td>
</tr>
<tr>
<td>10</td>
<td>1,955,007</td>
</tr>
<tr>
<td>11</td>
<td>518,410</td>
</tr>
<tr>
<td>12</td>
<td>149,945</td>
</tr>
<tr>
<td>13</td>
<td>44,616</td>
</tr>
<tr>
<td>14</td>
<td>13,740</td>
</tr>
<tr>
<td>15</td>
<td>4,476</td>
</tr>
<tr>
<td>16</td>
<td>1,542</td>
</tr>
<tr>
<td>17</td>
<td>536</td>
</tr>
<tr>
<td>18</td>
<td>167</td>
</tr>
<tr>
<td>19</td>
<td>71</td>
</tr>
<tr>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
</tbody>
</table>

# nodes as we do BFS out of a random node
MSN: Key Network Properties

Degree distribution: Heavily skewed
avg. degree = 14.4

Path length: 6.6

Clustering coefficient: 0.11

Are these values “expected”? Are they “surprising”? To answer this we need a null-model!
Erdös-Renyi Random Graph Model
What kinds of networks does such model produce?
Random Graph Model

- $n$ and $p$ do not uniquely determine the graph!
  - The graph is a result of a random process
- We can have many different realizations given the same $n$ and $p$

![Graph Examples]

$n = 10$
$p = 1/6$
How 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 in an undirected graph of \( n \) nodes.

\( P(E) \) is exactly the Binomial distribution

Number of successes in a sequence of \( E_{\text{max}} \) independent yes/no experiments.
What is expected degree of a node?

- Let $X_v$ be a rnd. var. measuring the degree of node $v$
- We want to know: $E[X_v] = \sum_{j=0}^{n-1} j \cdot P(X_v = j)$
  - For the calculation we will need: Linearity of expectation
    - For any random variables $Y_1, Y_2, \ldots, Y_k$
    - If $Y = Y_1 + Y_2 + \ldots Y_k$, then $E[Y] = \sum_i E[Y_i]$

An easier way:
- Decompose $X_v$ to $X_v = X_{v,1} + X_{v,2} + \ldots + X_{v,n-1}$
  - where $X_{v,u}$ is a $\{0, 1\}$-random variable which tells if edge $(v,u)$ exists or not

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

How to think about this?
- Prob. of node $u$ linking to node $v$ is $p$
- $u$ can link (flips a coin) to all other $(n-1)$ nodes
- Thus, the expected degree of node $u$ is: $p(n-1)$
Properties of $G_{np}$

Degree distribution: $P(k)$

Path length: $h$

Clustering coefficient: $C$

What are values of these properties for $G_{np}$?
Fact: 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 out of $n-1$:

- Probability of having $k$ edges
- Probability of missing the rest of the $n-1-k$ edges

Mean, variance of a binomial distribution

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

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

By the law of large numbers, 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 $k$. 

$$\sigma_k = \left[ \frac{1 - p}{p} \cdot \frac{1}{n-1} \right]^{1/2} \approx \frac{1}{(n-1)^{1/2}}$$
Clustering Coefficient of $G_{np}$

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

- **Edges in $G_{np}$ appear i.i.d. with prob. $p$**

- **So:** $e_i = p \frac{k_i(k_i - 1)}{2}$
  - Each pair is connected with prob. $p$
  - Number of distinct pairs of neighbors of node $i$ of degree $k_i$

- **Then:** $C = \frac{p \cdot k_i(k_i - 1)}{k_i(k_i - 1)} = \frac{\bar{k}}{n - 1} \approx \frac{\bar{k}}{n}$
  - Clustering coefficient of a random graph is small.
  - For a fixed avg. degree (that is $p=1/n$), $C$ decreases with the graph size $n$. 

Where $e_i$ is the number of edges between $i$'s neighbors.
Network Properties of $G_{np}$

Degree distribution:

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

Clustering coefficient:

$$C = p = \bar{k}/n$$

Path length: next!
To prove the diameter of a $G_{np}$ we define few concepts

**Define: Random k-Regular graph**

- Assume each node has $k$ spokes (half-edges)
  - $k=1$: Graph is a set of pairs
  - $k=2$: Graph is a set of cycles
  - $k=3$: Arbitrarily complicated graphs

Randomly pair them up!
Def: Expansion

- Graph $G(V, E)$ has expansion $\alpha$: if $\forall S \subseteq V$:
  
  $\# \text{ of edges leaving } S \geq \alpha \cdot \min(|S|, |V\setminus S|)$

- Or equivalently:
  
  $$\alpha = \min_{S \subseteq V} \frac{\# \text{ edges leaving } S}{\min(|S|, |V\setminus S|)}$$
Expansion: Intuition

\[ \alpha = \min_{S \subseteq V} \frac{\# \text{edges leaving } S}{\min(|S|, |V \setminus S|)} \]

(A big) graph with “good” expansion
Expansion is measure of robustness:
  - To disconnect \( l \) nodes, we need to cut \( \geq \alpha \cdot l \) edges

Low expansion:

High expansion:

Social networks:
  - “Communities”

\[
\alpha = \min_{S \subseteq V} \frac{\text{edges leaving } S}{\min(|S|, |V \setminus S|)}
\]
Expansion: k-Regular Graphs

- **k-regular graph** (every node has degree \( k \)):
  - Expansion is at most \( k \) (when \( S \) is a single node)

- Is there a graph on \( n \) nodes (\( n \to \infty \)), of fixed max deg. \( k \), so that expansion \( \alpha \) remains const?

**Examples:**

- **n\times n** grid: \( k = 4 \): \( \alpha = 2n/(n^2/4) \to 0 \)
  (\( S = n/2 \times n/2 \) square in the center)

- **Complete binary tree:**
  \( \alpha \to 0 \) for \( |S| = (n/2) - 1 \)

- **Fact:** For a random **3-regular graph** on \( n \) nodes, there is some const \( \alpha \) (\( \alpha > 0 \), independent of \( n \)) such that w.h.p. the expansion of the graph is \( \geq \alpha \)
Fact: In a graph on $n$ nodes with expansion $\alpha$ for all pairs of nodes $s$ and $t$ there is a path of $O((\log n) / \alpha)$ edges connecting them.

Proof:

- Proof strategy:
  - We want to show that from any node $s$ there is a path of length $O((\log n) / \alpha)$ to any other node $t$
  - Let $S_j$ be a set of all nodes found within $j$ steps of BFS from $s$.
  - How does $S_j$ increase as a function of $j$?
Proof (continued):

- Let $S_j$ be a set of all nodes found within $j$ steps of BFS from $s$.

- We want to relate $S_j$ and $S_{j+1}$.

\[
|S_{j+1}| \geq |S_j| + \frac{\alpha |S_j|}{k} =
\]

\[
|S_{j+1}| \geq |S_j| \left(1 + \frac{\alpha}{k}\right) = \left(1 + \frac{\alpha}{k}\right)^{j+1}
\]

At most $k$ edges “collide” at a node.

At least $\alpha |S_j|$ edges

Each of degree $k$
Proof (continued):

- In how many steps of BFS do we reach \( >n/2 \) nodes?
- Need \( j \) so that: \( S_j = \left(1 + \frac{\alpha}{k}\right)^j \geq \frac{n}{2} \)
- Let’s set: \( j = \frac{k \log_2 n}{\alpha} \)
- Then:
  \[
  \left(1 + \frac{\alpha}{k}\right)^{\frac{k \log_2 n}{\alpha}} \geq 2^{\log_2 n} = n > \frac{n}{2}
  \]
- In \( 2k/\alpha \cdot \log n \) steps \( |S_j| \) grows to \( \Theta(n) \).
  So, the diameter of \( G \) is \( O(\log(n)/\alpha) \)

\[
e = \lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x
\]
Network Properties of $G_{np}$

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

Path length: $O(\log n)$

Clustering coefficient: $C = p = \bar{k} / n$
Degree distribution:

Path length: 6.6

Clustering coefficient: 0.11

$O(\log n)$

$\approx 8.2$

$\bar{k}/n$

$\approx 8 \cdot 10^{-8}$
Real Networks vs. $G_{np}$

- **Are real networks like random graphs?**
  - Giant connected component: 😊
  - Average path length: 😊
  - Clustering Coefficient: 😞
  - Degree Distribution: 😞

- **Problems with the random networks model:**
  - Degree distribution differs from that of real networks
  - Giant component in most real network does NOT emerge through a phase transition
  - No local structure – clustering coefficient is too low

- **Most important: Are real networks random?**
  - The answer is simply: NO!
If $G_{np}$ is wrong, why did we spend time on it?

- It is the reference model for the rest of the class.
- It will help us calculate many quantities, that can then be compared to the real data.
- It will help us understand to what degree is a particular property the result of some random process.

So, while $G_{np}$ is WRONG, it will turn out to be extremely USEFUL!
EXTRA: “Evolution” of the $G_{np}$

What happens to $G_{np}$ when we vary $p$?
Remember, expected degree $E[X_v] = (n - 1) p$

We want $E[X_v]$ to be independent of $n$

So let: $p = c/(n-1)$

Observation: If we build random graph $G_{np}$ with $p = c/(n-1)$ we have many isolated nodes

Why?

$P[\nu \text{ has degree 0}] = (1 - p)^{n-1} = \left(1 - \frac{c}{n - 1}\right)^{n-1} \xrightarrow{n \to \infty} e^{-c}$

By definition:

$$e = \lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x$$

Use substitution $\frac{1}{x} = \frac{c}{n-1}$

$$e$$
How big do we have to make $p$ before we are likely to have no isolated nodes?

We know: $P[v \text{ has degree } 0] = e^{-c}$

Event we are asking about is:

- $I = \text{some node is isolated}$
- $I = \bigcup_{v \in N} I_v$ where $I_v$ is the event that $v$ is isolated

We have:

$$P(I) = P\left( \bigcup_{v \in N} I_v \right) \leq \sum_{v \in N} P(I_v) = ne^{-c}$$

Union bound

$$\left| \bigcup_{i} A_i \right| \leq \sum_{i} |A_i|$$
We just learned: $P(I) = n e^{-c}$

Let’s try:

- $c = \ln n$ then: $n e^{-c} = n e^{-\ln n} = n \cdot 1/n = 1$
- $c = 2 \ln n$ then: $n e^{-2 \ln n} = n \cdot 1/n^2 = 1/n$

So if:

- $p = \ln n$ then: $P(I) = 1$
- $p = 2 \ln n$ then: $P(I) = 1/n \to 0$ as $n \to \infty$
Graph structure of $G_{np}$ as $p$ changes:

- Emergence of a Giant Component:
  - Avg. degree $k = 2E/n$ or $p = k/(n-1)$
    - $k = 1-\varepsilon$: all components are of size $\Omega(\log n)$
    - $k = 1+\varepsilon$: 1 component of size $\Omega(n)$, others have size $\Omega(\log n)$
G_{np} Simulation Experiment

![Graph showing the fraction of nodes in the largest component against p*(n-1)]

- G_{np}, n=100k, p(n-1) = 0.5 ... 3