Chapter 4

Link Analysis & Authority Ranking

Sergej Sizov
Information Retrieval
Summer term 2009
Link Analysis: Early Approaches

Basic Assumptions

- Hyperlinks contain information about the human judgment of a site
- The more incoming links to a site, the more it is judged important

(Bray 1996)

- The visibility of a site is measured by the number of other sites pointing to it
- The luminosity of a site is measured by the number of other sites to which it points
  → Limitation: failure to capture the relative importance of different parents (children) sites
Early Approaches (2)

(Mark 1988)

- To calculate the score $S$ of a document at vertex $v$

$$S(v) = s(v) + \frac{1}{|ch[v]|} \sum_{w \in |ch(v)|} S(w)$$

$v$: a vertex in the hypertext graph $G = (V, E)$
$S(v)$: the global score
$s(v)$: the score if the document is isolated
$ch(v)$: children of the document at vertex $v$

- Limitation:
  - Require $G$ to be a directed acyclic graph (DAG)
  - If $v$ has a single link to $w$, $S(v) > S(w)$
  - If $v$ has a long path to $w$ and $s(v) < s(w)$, then $S(v) > S(w)$
  - unreasonable
Early Approaches (3)

(Marchiori 1997)

- Hyper information should complement textual information to obtain the overall information

\[ S(v) = s(v) + h(v) \]

- \( S(v) \): overall information
- \( s(v) \): textual information
- \( h(v) \): hyper information

\[ h(v) = \sum_{w \in \text{ch}[v]} F^{r(v, w)} S(w) \]

- \( F \): a fading constant, \( F \in (0, 1) \)
- \( r(v, w) \): the rank of \( w \) after sorting the children of \( v \) by \( S(w) \)

\[ \rightarrow \text{a remedy of the previous approach (Mark 1988)} \]
Improving Precision by Authority Scores

Goal:
Higher ranking of URLs with high authority regarding volume, significance, freshness, authenticity of content → improve precision of search results

Approaches (all interpreting the Web as a directed graph $G$):
• citation or impact rank $(q) \sim$ indegree $(q)$
• PageRank (by Lawrence Page)
• HITS algorithm (by Jon Kleinberg)

Combining relevance and authority ranking:
• by weighted sum with appropriate coefficients (Google)
• by initial relevance ranking and iterative improvement via authority ranking (HITS)
Web Structure: Small Diameter

Small World Phenomenon (Milgram 1967)
Studies on Internet Connectivity (1999)

suggested small world phenomenon: low-diameter graph
\[
\text{(diameter} = \max \{\text{shortest path} (x,y) \mid \text{nodes} x \text{ and } y\})
\]

Source: Bill Cheswick and Hal Burch,
http://research.lumeta.com/ches/map/index.html

Source: KC Claffy,
http://www.caida.org/outreach/papers/1999/Nae/Nae.html
Web Structure: Connected Components

Study of Web Graph (Broder et al. 2000)

SCC = strongly connected component

..strongly connected core tends to have small diameter

Source: A.Z. Broder et al., WWW 2000
Web Structure: Power-Law Degrees

Study of Web Graph (Broder et al. 2000)

- power-law distributed degrees: $P[\text{degree}=k] \sim (1/k)^\alpha$
  with $\alpha \approx 2.1$ for indegrees and $\alpha \approx 2.7$ for outdegrees
The Random Surfer Model

Idea: incoming links are endorsements & increase page authority, authority is higher if links come from high-authority pages

\[ PR(q) = \sum_{p \in IN(q)} PR(p) \times t(p, q) \]

Authority (page q) = stationary prob. of visiting q

random walk: uniformly random choice of links + random jumps
A stochastic matrix is an $n \times n$ matrix $M$
with row sum $\Sigma_{j=1..n} M_{ij} = 1$ for each row $i$

Random surfer follows a stochastic matrix

Theorem:
For every stochastic matrix $M$
all Eigenvalues $\lambda$ have the property $|\lambda| \leq 1$
and there is an Eigenvector $x$ with Eigenvalue $1$ s.t. $x \geq 0$ and $\|x\|_1 = 1$

Suggests power iteration $x^{(i+1)} = M^T x^{(i)}$

But: real Web graph
has sinks, may be periodic, is not strongly connected
Random Surfer: “Rank Sink” Problem

“Rank Sink” Problem
- In general, many Web pages have no inlinks/outlinks
- It results in dangling edges in the graph

E.g.
- no parent $\rightarrow$ rank 0
  $M^T$ converges to a matrix
  whose last column is all zero
- no children $\rightarrow$ no solution
  $M^T$ converges to zero matrix
Google’s PageRank

Idea: incoming links are endorsements & increase page authority, authority is higher if links come from high-authority pages

\[ PR(q) = \varepsilon \cdot j(q) + (1 - \varepsilon) \cdot \sum_{p \in IN(q)} PR(p) \cdot t(p, q) \]

Authority (page q) = stationary prob. of visiting q

random walk: uniformly random choice of links + random jumps
**Markov Chains**

- State set: finite or infinite
- Time: discrete or continuous
- State transition probabilities: $p_{ij}$
- State probabilities in step $t$: $p_i^{(t)} = P[S(t)=i]$

Markov property: $P[S(t)=i \mid S(0), \ldots, S(t-1)] = P[S(t)=i \mid S(t-1)]$

Interested in stationary state probabilities:

$$p_j := \lim_{t \to \infty} p_j^{(t)} = \lim_{t \to \infty} \sum_k p_k^{(t-1)} p_{kj}$$

$$p_j = \sum_k p_k p_{kj}$$

$$\sum_j p_j = 1$$

$p_0 = 0.8 p_0 + 0.5 p_1 + 0.4 p_2$

$p_1 = 0.2 p_0 + 0.3 p_2$

$p_2 = 0.5 p_1 + 0.3 p_2$

$p_0 + p_1 + p_2 = 1$

$\Rightarrow p_0 \approx 0.657, p_1 = 0.2, p_2 \approx 0.143$
Markov Chains: Formal Definition

A **stochastic process** is a family of random variables \( \{X(t) \mid t \in T\} \). 
T is called parameter space, and the domain M of \( X(t) \) is called state space. T and M can be discrete or continuous.

A stochastic process is called **Markov process** if for every choice of \( t_1, ..., t_{n+1} \) from the parameter space and every choice of \( x_1, ..., x_{n+1} \) from the state space the following holds:

\[
P \left[ X(t_{n+1}) = x_{n+1} \mid X(t_1) = x_1 \land X(t_2) = x_2 \land ... \land X(t_n) = x_n \right]
= P \left[ X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n \right]
\]

A Markov process with discrete state space is called **Markov chain**. A canonical choice of the state space are the natural numbers. Notation for Markov chains with discrete parameter space: \( X_n \) rather than \( X(t_n) \) with \( n = 0, 1, 2, ... \)
Properties of Markov Chains with Discrete Parameter Space (1)

The Markov chain $X_n$ with discrete parameter space is

**homogeneous** if the transition probabilities

$p_{ij} := P[X_{n+1} = j \mid X_n = i]$ are independent of $n$

**irreducible** if every state is reachable from every other state with positive probability:

$$\sum_{n=1}^{\infty} P[X_n = j \mid X_0 = i] > 0 \quad \text{for all } i, j$$

**aperiodic** if every state $i$ has period 1, where the period of $i$ is the greatest common divisor of all (recurrence) values $n$ for which

$$P[X_n = i \land X_k \neq i \text{ for } k = 1, ..., n-1 \mid X_0 = i] > 0$$
The Markov chain $X_n$ with discrete parameter space is **positive recurrent** if for every state $i$ the recurrence probability is 1 and the mean recurrence time is finite:

$$\sum_{n=1}^{\infty} P[ X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 \mid X_0 = i ] = 1$$

and

$$\sum_{n=1}^{\infty} n P[ X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 \mid X_0 = i ] < \infty$$

**ergodic** if it is homogeneous, irreducible, aperiodic, and positive recurrent.
Results on Markov Chains with Discrete Parameter Space (1)

For the **n-step transition probabilities**

\[ p^{(n)}_{ij} := P [ \ X_n = j \mid X_0 = i ] \]  

the following holds:

\[
p^{(n)}_{ij} = \sum_k p^{(n-1)}_{ik} p_{kj} \quad \text{with} \quad p^{(1)}_{ij} := p_{ik}
\]

\[
= \sum_k p^{(n-l)}_{ik} p^{(l)}_{kj} \quad \text{for} \ 1 \leq l \leq n - 1
\]

in matrix notation:

\[ P^{(n)} = P^n \]

For the **state probabilities after n steps**

\[ \pi^{(n)}_j := P [ \ X_n = j ] \]  

the following holds:

\[
\pi^{(n)}_j = \sum_i \pi^{(0)}_i p^{(n)}_{ij} \quad \text{with initial state probabilities} \quad \pi^{(0)}_i
\]

in matrix notation:

\[ \Pi^{(n)} = \Pi^{(0)} P^{(n)} \ (Chapman-Kolmogorov \ equation) \]
Every homogeneous, irreducible, aperiodic Markov chain with a finite number of states is positive recurrent and ergodic.

For every ergodic Markov chain there exist stationary state probabilities

These are independent of \( \Pi^{(0)} \)

and are the solutions of the following system of linear equations:

\[
\pi_j = \sum_i \pi_i p_{ij} \quad \text{for all } j
\]

\[
\sum_j \pi_j = 1
\]

in matrix notation:

\[
\Pi = \Pi P
\]

(\text{balance equations})

\[
\Pi \vec{1} = 1
\]
Page Rank as a Markov Chain Model

Model a random walk of a Web surfer as follows:
- follow outgoing hyperlinks with uniform probabilities
- perform „random jump“ with probability $\varepsilon$

→ ergodic Markov chain

The PageRank of a URL is the stationary visiting probability of URL in the above Markov chain.

Further generalizations have been studied (e.g. random walk with back button etc.)

Drawback of Page rank method:
Page rank is query-independent and orthogonal to relevance
Example: Page Rank Computation

\[ P = \begin{pmatrix} 0.0 & 0.5 & 0.5 \\ 0.1 & 0.0 & 0.9 \\ 0.9 & 0.1 & 0.0 \end{pmatrix} \]

\[ \varepsilon = 0.2 \]

\[ \Phi \approx \begin{pmatrix} 0.333 \\ 0.333 \\ 0.333 \end{pmatrix} \]

\[ \Pi^{(1)} \approx \begin{pmatrix} 0.333 \\ 0.200 \\ 0.466 \end{pmatrix} \]

\[ \Pi^{(2)} \approx \begin{pmatrix} 0.439 \\ 0.212 \\ 0.346 \end{pmatrix} \]

\[ \Pi^{(3)} \approx \begin{pmatrix} 0.332 \\ 0.253 \\ 0.401 \end{pmatrix} \]

\[ \Pi^{(4)} \approx \begin{pmatrix} 0.385 \\ 0.176 \\ 0.527 \end{pmatrix} \]

\[ \Pi^{(5)} \approx \begin{pmatrix} 0.491 \\ 0.244 \\ 0.350 \end{pmatrix} \]

\[ \pi_1 = 0.1 \pi_2 + 0.9 \pi_3 \]
\[ \pi_2 = 0.5 \pi_1 + 0.1 \pi_3 \]
\[ \pi_3 = 0.5 \pi_1 + 0.9 \pi_2 \]
\[ \pi_1 + \pi_2 + \pi_3 = 1 \]

\[ \Rightarrow \pi_1 \approx 0.3776, \pi_2 \approx 0.2282, \pi_3 \approx 0.3942 \]
HITS Algorithm: Hyperlink-Induced Topic Search (1)

Idea:
Determine
• good content sources: **Authorities** (high indegree)
• good link sources: **Hubs** (high outdegree)

Find
• better authorities that have good hubs as predecessors
• better hubs that have good authorities as successors

For Web graph $G=(V,E)$ define for nodes $p, q \in V$

**authority score**
$$x_q = \sum_{(p,q) \in E} y_p$$

and

**hub score**
$$y_p = \sum_{(p,q) \in E} x_q$$
HITS Algorithm (2)

Authority and hub scores in matrix notation:

\[ \tilde{x} = A^T \tilde{y} \quad \tilde{y} = A \tilde{x} \]

Iteration with adjacency matrix \( A \):

\[ \tilde{x} := A^T \tilde{y} := A^T A \tilde{x} \quad \tilde{y} := A \tilde{x} := A A^T \tilde{y} \]

Using Linear Algebra, we can prove:
- \( x \) and \( y \) converge
- \( x \) and \( y \) are \textbf{Eigenvectors} of \( A^T A \) and \( AA^T \), resp.

Intuitive interpretation:

\( M^{(auth)} := A^T A \) is the cocitation matrix: \( M^{(auth)}_{ij} \) is the number of nodes that point to both \( i \) and \( j \)

\( M^{(hub)} := AA^T \) is the bibliographic-coupling matrix: \( M^{(hub)}_{ij} \) is the number of nodes to which both \( i \) and \( j \) point
Implementation of the HITS Algorithm

1) Determine sufficient number (e.g. 50-200) of „root pages“ via relevance ranking (e.g. using \( tf*idf \) ranking)
2) Add all successors of root pages
3) For each root page add up to \( d \) predecessors
4) Compute iteratively
   the authority and hub scores of this „base set“ (of typically 1000-5000 pages)
   with initialization \( x_q := y_p := 1 / |\text{base set}| \)
   and normalization after each iteration
   \( \rightarrow \) converges to principal Eigenvector (Eigenvector with largest Eigenvalue (in the case of multiplicity 1))
5) Return pages in descending order of authority scores (e.g. the 10 largest elements of vector \( x \))

Drawback of HITS algorithm:
relevance ranking within root set is not considered
Example: HITS Algorithm

root set

base set
(Bharat and Henzinger, 1998)

- HITS problems
  1) The document can contain many *identical* links to the same document in another host
  2) Links are generated automatically (e.g. messages posted on newsgroups)

- Solutions
  1) Assign weight to *identical* multiple edges, which are inversely proportional to their multiplicity
  2) Prune irrelevant nodes or regulating the influence of a node with a relevance weight
**Improved HITS Algorithm**

Potential weakness of the HITS algorithm:
- irritating links (automatically generated links, spam, etc.)
- topic drift (e.g. from „Jaguar car“ to „car“ in general)

Improvement:
- Introduce **edge weights**:
  - 0 for links within the same host,
  - 1/k with k links from k URLs of the same host to 1 URL (xweight)
  - 1/m with m links from 1 URL to m URLs on the same host (yweight)
- Consider **relevance weights** w.r.t. query topic (e.g. tf*idf)

→ Iterative computation of

**authority score**  \[ x_q = \sum_{(p,q)\in E} y_p \times \text{topic score}(p) \times \text{xweight}(p,q) \]

**hub score**  \[ y_p = \sum_{(p,q)\in E} x_q \times \text{topic score}(q) \times \text{yweight}(p,q) \]
SALSA

SALSA (Lempel, Moran 2001)

- Probabilistic extension of the HITS algorithm
- Random walk is carried out by following hyperlinks both in the forward and in the backward direction

Two separate random walks

- Hub walk
- Authority walk
Forming a Bipartite Graph in SALSA

\[ G \]

\[ U \]
Random Walks

Hub walk

- Follow a Web link from a page $u_h$ to a page $w_a$ (a forward link) and then
- Immediately traverse a backlink going from $w_a$ to $v_h$, where $(u,w) \in E$ and $(v,w) \in E$

Authority Walk

- Follow a Web link from a page $w(a)$ to a page $u(h)$ (a backward link) and then
- Immediately traverse a forward link going back from $v_h$ to $w_a$ where $(u,w) \in E$ and $(v,w) \in E$

Lempel and Moran (2001) showed theoretically that SALSA weights are more robust than HITS weights in the presence of the Tightly Knit Community (TKC) Effect.

- This effect occurs when a small collection of pages (related to a given topic) is connected so that every hub links to every authority and includes as a special case the mutual reinforcement effect

The pages in a community connected in this way can be ranked highly by HITS, higher than pages in a much larger collection where only some hubs link to some authorities
SALSA: Random Walk on Hubs and Authorities

View each node $v$ of the link graph as two nodes $v_h$ and $v_a$

Construct bipartite undirected graph $G'(V',E')$ from link graph $G(V,E)$:
$V' = \{v_h | v \in V \text{ and outdegree}(v)>0\} \cup \{v_a | v \in V \text{ and indegree}(v)>0\}$
$E' = \{(v_h, w_a) | (v, w) \in E\}$

Stochastic hub matrix $H$:
$$h_{ij} = \sum_k \frac{1}{\text{deg}ree(i_h)} \frac{1}{\text{deg}ree(k_a)}$$
for hubs $i$, $j$ and $k$ ranging over all nodes with $(i_h, k_a), (k_a, j_h) \in E'$

Stochastic authority matrix $A$:
$$a_{ij} = \sum_k \frac{1}{\text{deg}ree(i_a)} \frac{1}{\text{deg}ree(k_h)}$$
for authorities $i$, $j$ and $k$ ranging over all nodes with $(i_a, k_h), (k_h, j_a) \in E'$

The corresponding Markov chains are ergodic on connected component

The stationary solutions for these Markov chains are:
$\pi[v_h] \sim \text{outdegree}(v)$ for $H$ and $\pi[v_a] \sim \text{indegree}(v)$ for $A$
Comparison and Extensions

Literature contains plethora of variations on Page-Rank and HITS

Key points are:
• mutual reinforcement between hubs and authorities
• re-scale edge weights (normalization)

Unified notation (for link graph with n nodes):

- \( L \) - \( n \times n \) link matrix, \( L_{ij} = 1 \) if there is an edge \((i,j)\), 0 else
- \( \text{din} \) - \( n \times 1 \) vector with \( \text{din}_i = \text{indegree}(i) \), \( \text{Din}_{n \times n} = \text{diag}(\text{din}) \)
- \( \text{dout} \) - \( n \times 1 \) vector with \( \text{dout}_i = \text{outdegree}(i) \), \( \text{Dout}_{n \times n} = \text{diag}(\text{dout}) \)
- \( x \) - \( n \times 1 \) authority vector
- \( y \) - \( n \times 1 \) hub vector
- \( \text{Iop} \) - operation applied to incoming links
- \( \text{Oop} \) - operation applied to outgoing links
**HITS**: \( x = \text{Iop}(y), \ y = \text{Oop}(x) \) with \( \text{Iop}(y) = L^T y \), \( \text{Oop}(x) = Lx \)

**PageRank**: \( x = \text{Iop}(x) \) with \( \text{Iop}(x) = P^T x \) with \( P^T = L^T D_{out}^{-1} \)

or \( P^T = \alpha L^T D_{out}^{-1} + (1-\alpha) \frac{1}{n} e e^T \)

**SALSA** (PageRank-style computation with mutual reinforcement):
\( x = \text{Iop}(y) \) with \( \text{Iop}(y) = P^T y \) with \( P^T = L^T D_{out}^{-1} \)
\( y = \text{Oop}(x) \) with \( \text{Oop}(x) = Q x \) with \( Q = L D_{in}^{-1} \)

and other models of link analysis can be cast into this framework, too
Link Analysis: Stability

Whether the link analysis algorithms based on eigenvectors are stable in the sense that results don’t change significantly?

The connectivity of a portion of the graph is changed arbitrary
  - How will it affect the results of algorithms?
Stability of HITS

(Ng et al, 2001)

- A bound on the number of hyperlinks $k$ that can added or deleted from one page without affecting the authority or hubness weights
- It is possible to perturb a symmetric matrix by a quantity that grows as $\delta$ that produces a constant perturbation of the dominant eigenvector

\[
\| \vec{a} - \vec{a}_* \|_2 \leq \alpha \\

k \leq \left( \sqrt{d + \frac{\alpha \delta}{4 + \sqrt{2\alpha}}} - \sqrt{d} \right)^2
\]

$\delta$: eigengap $\lambda_1 - \lambda_2$

$d$: maximum outdegree of $G$
Stability of PageRank

Ng et al (2001)

$$\|\vec{r} - \vec{r}_*\| \geq \frac{2 \sum_{j \in V^*} r(j)}{\varepsilon}$$

$V^*$: the set of vertices touched by the perturbation

The parameter $\varepsilon$ of the mixture model has a stabilization role.

If the set of pages affected by the perturbation have a small rank, the overall change will also be small.

$$\|\vec{r} - \vec{r}_*\| \geq \frac{1 - \varepsilon}{\varepsilon} \sum_{j \in V^*} \delta(j) r(j)$$

tighter bound by Bianchini et al (2001)

$\delta(j) \geq 2$ depends on the edges incident on $j$
Link Analysis: Advanced Issues

- Topic-specific authority ranking
- Personalized PageRank
  - Personalization vectors
  - Query logs
  - Click streams
- Efficiency issues
- Online Page Importance (OPIC)
- Time-aware link analysis
Web Models: Understanding Large Graphs

What are the statistics of real life networks?
Can we explain how the networks were generated?
Small World Experiment (Six degrees of separation)

Stanley Milgram: 1967: Letters were handed out to people in Nebraska to be sent to a target (stock broker) in Boston
- People were instructed to pass on the letters to someone they knew on first-name basis
- The letters that reached the destination followed paths of avg length 5.2 (i.e. around 6)

Duncan Watts: 2001: Milgram's experiment recreated on the internet
- using an e-mail message as the "package" that needed to be delivered, with 48,000 senders and 19 targets (in 157 countries).
- the avg number of intermediaries was also around 6.

See also:
- The Kevin Bacon game
- The Erdös number
- etc.
Kevin Bacon Experiment
Craig Fass, Brian Turtle and Mike Ginelli: 1994: motivated by Bacon's most recent movie „The Air Up There” and his career discussion
Vertices: actors and actresses
Edge between u and v if they appeared in a movie together

Is Kevin Bacon the most connected actor?

See also:
http://oracleofbacon.org/

Kevin Bacon

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Average distance</th>
<th># of movies</th>
<th># of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rod Steiger</td>
<td>2.537527</td>
<td>112</td>
<td>2562</td>
</tr>
<tr>
<td>2</td>
<td>Donald Pleasence</td>
<td>2.542376</td>
<td>180</td>
<td>2874</td>
</tr>
<tr>
<td>3</td>
<td>Martin Sheen</td>
<td>2.551210</td>
<td>136</td>
<td>3501</td>
</tr>
<tr>
<td>4</td>
<td>Christopher Lee</td>
<td>2.552497</td>
<td>201</td>
<td>2993</td>
</tr>
<tr>
<td>5</td>
<td>Robert Mitchum</td>
<td>2.557181</td>
<td>136</td>
<td>2905</td>
</tr>
<tr>
<td>6</td>
<td>Charlton Heston</td>
<td>2.566284</td>
<td>104</td>
<td>2552</td>
</tr>
<tr>
<td>7</td>
<td>Eddie Albert</td>
<td>2.567036</td>
<td>112</td>
<td>3333</td>
</tr>
<tr>
<td>8</td>
<td>Robert Vaughn</td>
<td>2.570193</td>
<td>126</td>
<td>2761</td>
</tr>
<tr>
<td>9</td>
<td>Donald Sutherland</td>
<td>2.577880</td>
<td>107</td>
<td>2865</td>
</tr>
<tr>
<td>10</td>
<td>John Gielgud</td>
<td>2.578980</td>
<td>122</td>
<td>2942</td>
</tr>
<tr>
<td>11</td>
<td>Anthony Quinn</td>
<td>2.579750</td>
<td>146</td>
<td>2978</td>
</tr>
<tr>
<td>12</td>
<td>James Earl Jones</td>
<td>2.584440</td>
<td>112</td>
<td>3787</td>
</tr>
</tbody>
</table>

Kevin Bacon
No. of movies : 46
No. of actors : 1811
Average separation: 2.79
Quantities of Interest

Connected components:
- how many, and how large?

Network diameter:
- maximum (worst-case) or average?
- exclude infinite distances? (disconnected components)
- the small-world phenomenon

Clustering:
- to what extent that links tend to cluster “locally”?
- what is the balance between local and long-distance connections?
- what roles do the two types of links play?

Degree distribution:
- what is the typical degree in the network?
- what is the overall distribution?
Graph theory: undirected graph notation

Graph $G=(V,E)$
- $V =$ set of vertices
- $E =$ set of edges

undirected graph
$E=\{(1,2),(1,3),(2,3),(3,4),(4,5)\}$
Graph theory: directed graph notation

Graph \( G=(V,E) \)
- \( V \) = set of vertices
- \( E \) = set of edges

directed graph
\( E=\{\langle 1,2 \rangle, \langle 2,1 \rangle <1,3>, \langle 3,2 \rangle, \langle 3,4 \rangle, \langle 4,5 \rangle\} \)
undirected graph: degree distribution

degree $d(i)$ of node $i$
  - number of edges incident on node $i$

degree sequence
  - $[d(1), d(2), d(3), d(4), d(5)]$
  - $[2, 2, 3, 2, 1]$

degree distribution
  - $[(1, 1), (2, 3), (3, 1)]$
Directed graph: in/outdegrees

in-degree $d_{in}(i)$ of node $i$
  - number of edges pointing to node $i$

out-degree $d_{out}(i)$ of node $i$
  - number of edges leaving node $i$

in-degree sequence
  - [1,2,1,1,1]

out-degree sequence
  - [2,1,2,1,0]
Degree distributions

Problem: find the probability distribution that best fits the observed data

\[ f_k = \text{fraction of nodes with degree } k \]
\[ = \text{probability of a randomly selected node to have degree } k \]
Power-law distributions

The degree distributions of most real-life networks follow a power law

\[ p(k) = Ck^{-\alpha} \]

Right-skewed/Heavy-tail distribution
- There is a non-negligible fraction of nodes that has very high degree (hubs)
- Scale-free: no characteristic scale, average is not informative
- Highly concentrated around the mean
- The probability of very high degree nodes is exponentially small
Power-law signature

\[ p(k) = Ck^{-\alpha} \]

Power-law distribution gives a line in the log-log plot

\[ \log p(k) = -\alpha \log k + \log C \]

\[ \alpha : \text{power-law exponent (typically } 2 \leq \alpha \leq 3) \]
Power-law: Examples

Taken from [Newman 2003]
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>(\ell)</th>
<th>(\alpha)</th>
<th>(C^{(1)})</th>
<th>(C^{(2)})</th>
<th>(r)</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449 913</td>
<td>25 516 482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7 673</td>
<td>55 392</td>
<td>14.44</td>
<td>4.60</td>
<td>-</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253 339</td>
<td>496 489</td>
<td>3.92</td>
<td>7.57</td>
<td>-</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52 909</td>
<td>245 300</td>
<td>9.27</td>
<td>6.19</td>
<td>-</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1 520 251</td>
<td>11 803 064</td>
<td>15.53</td>
<td>4.92</td>
<td>-</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47 000 000</td>
<td>80 000 000</td>
<td>3.16</td>
<td>2.1</td>
<td>-</td>
<td>8, 9</td>
<td>136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59 912</td>
<td>86 300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16 881</td>
<td>57 029</td>
<td>3.38</td>
<td>5.22</td>
<td>-</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>-</td>
<td>0.005</td>
<td>0.001</td>
<td>0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2 810</td>
<td></td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265, 266</td>
</tr>
<tr>
<td>WWW nd. edu</td>
<td>directed</td>
<td>269 504</td>
<td>1 497 135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>-0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW AltaVista</td>
<td>directed</td>
<td>203 549 046</td>
<td>2 130 000 000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783 339</td>
<td>6 716 198</td>
<td>8.57</td>
<td>3.0/-</td>
<td></td>
<td></td>
<td>351</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1 022</td>
<td>5 103</td>
<td>4.99</td>
<td>4.87</td>
<td>-</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460 902</td>
<td>17 000 000</td>
<td>70.13</td>
<td>2.7</td>
<td></td>
<td>0.44</td>
<td>119, 157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10 697</td>
<td>31 992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.6</td>
<td>0.035</td>
<td>0.39</td>
<td>-0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4 941</td>
<td>6 594</td>
<td>2.67</td>
<td>18.99</td>
<td>-</td>
<td>0.10</td>
<td>0.080</td>
<td>-0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19 603</td>
<td>66.79</td>
<td>2.16</td>
<td>-</td>
<td>0.69</td>
<td>-0.033</td>
<td>366</td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1 439</td>
<td>1 723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>-0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1 377</td>
<td>2 213</td>
<td>1.61</td>
<td>1.51</td>
<td>-</td>
<td>0.033</td>
<td>0.012</td>
<td>-0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24 097</td>
<td>53 248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>-0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1 296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>-0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3 686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>-0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2 115</td>
<td>2 240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>-0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>-</td>
<td>0.16</td>
<td>0.23</td>
<td>-0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>-</td>
<td>0.20</td>
<td>0.087</td>
<td>-0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2 359</td>
<td>7.68</td>
<td>3.97</td>
<td>-</td>
<td>0.18</td>
<td>0.28</td>
<td>-0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

**TABLE II** Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices \(n\); total number of edges \(m\); mean degree \(z\); mean vertex–vertex distance \(\ell\); exponent \(\alpha\) of degree distribution if the distribution follows a power law (or “-” if not); in/out-degree exponents are given for directed graphs; clustering coefficient \(C^{(1)}\) from Eq. (3); clustering coefficient \(C^{(2)}\) from Eq. (6); and degree correlation coefficient \(r\), Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Web Structure: Power-Law Degrees

Study of Web Graph (Broder et al. 2000)

- power-law distributed degrees: $P[\text{degree}=k] \sim (1/k)^\alpha$
  with $\alpha \approx 2.1$ for indegrees and $\alpha \approx 2.7$ for outdegrees
Clustering (Transitivity) coefficient

Measures the density of triangles (local clusters) in the graph

Two different ways to measure it:

The ratio of the means

\[
C^{(1)} = \frac{\sum_{i} \text{triangles centered at node } i}{\sum_{i} \text{triples centered at node } i}
\]
Link Analysis and Authority Ranking

Example

\[ C^{(1)} = \frac{3}{1 + 1 + 6} = \frac{3}{8} \]
Clustering (Transitivity) coefficient

Clustering coefficient for node $i$

$$C_i = \frac{\text{triangles centered at node } i}{\text{triples centered at node } i}$$

The mean of the ratios

$$C^{(2)} = \frac{1}{n} C_i$$
Example

The two clustering coefficients give different measures:

\[ C^{(2)} = \frac{1}{5} \left(1 + 1 + \frac{1}{6}\right) = \frac{13}{30} \]

\[ C^{(1)} = \frac{3}{8} \]

The two clustering coefficients give different measures. \( C^{(2)} \) increases with nodes with low degree.
### Collective Statistics (M. Newman 2003)

<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>$n$</th>
<th>$m$</th>
<th>$z$</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s).</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449 913</td>
<td>25 516 482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7675</td>
<td>55 392</td>
<td>14.44</td>
<td>4.60</td>
<td>-</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253 339</td>
<td>496 489</td>
<td>3.92</td>
<td>7.57</td>
<td>-</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52 909</td>
<td>245 300</td>
<td>9.27</td>
<td>6.19</td>
<td>-</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1520 251</td>
<td>11 803 064</td>
<td>15.53</td>
<td>4.92</td>
<td>-</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47 000 000</td>
<td>80 000 000</td>
<td>3.16</td>
<td>2.1</td>
<td>-</td>
<td>0.005</td>
<td>0.001</td>
<td>-0.029</td>
<td>45</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59 912</td>
<td>86 300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16 881</td>
<td>57 029</td>
<td>3.38</td>
<td>5.22</td>
<td>-</td>
<td>0.17</td>
<td>0.13</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>575</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>-</td>
<td>0.005</td>
<td>0.001</td>
<td>-0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2810</td>
<td>3.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269 504</td>
<td>1 497 135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>-0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>203 549 046</td>
<td>2 130 000 000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.24</td>
<td>0.39</td>
<td>-0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783 339</td>
<td>6 716 198</td>
<td>8.57</td>
<td>-</td>
<td>3.0/-</td>
<td>0.35</td>
<td>0.44</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1022</td>
<td>5 103</td>
<td>4.99</td>
<td>4.87</td>
<td>-</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460 902</td>
<td>17 000 000</td>
<td>70.13</td>
<td>-</td>
<td>2.7</td>
<td>0.44</td>
<td>0.44</td>
<td>-</td>
<td>119, 157</td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10 697</td>
<td>31 992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>-0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4941</td>
<td>6 594</td>
<td>2.67</td>
<td>18.99</td>
<td>-</td>
<td>0.10</td>
<td>0.080</td>
<td>-0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19 603</td>
<td>66.79</td>
<td>2.16</td>
<td>-</td>
<td>0.09</td>
<td>0.033</td>
<td>0.366</td>
<td>-</td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1439</td>
<td>1 723</td>
<td>1.20</td>
<td>2.24</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>-0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1377</td>
<td>2 213</td>
<td>1.61</td>
<td>1.51</td>
<td>-</td>
<td>0.033</td>
<td>0.012</td>
<td>-0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24 977</td>
<td>53 248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>-0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1 296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>-0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3 686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>-0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2 115</td>
<td>2 240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>-0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>-</td>
<td>0.16</td>
<td>0.23</td>
<td>-0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>-</td>
<td>0.20</td>
<td>0.087</td>
<td>-0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2 359</td>
<td>7.68</td>
<td>3.97</td>
<td>-</td>
<td>0.18</td>
<td>0.28</td>
<td>-0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

**TABLE II** Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices $n$; total number of edges $m$; mean degree $z$; mean vertex-vertex distance $\ell$; exponent $\alpha$ of degree distribution if the distribution follows a power law (or “-$\alpha$” if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient $r$, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Clustering coefficient for random graphs

The probability of two of your neighbors also being neighbors is $p$, independent of local structure

- clustering coefficient $C = p$
- when $z$ is fixed $C = z/n = O(1/n)$

<table>
<thead>
<tr>
<th>Network</th>
<th>$n$</th>
<th>$z$</th>
<th>$C$ measured</th>
<th>$C$ for random graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet [153]</td>
<td>6,374</td>
<td>3.8</td>
<td>0.24</td>
<td>0.00060</td>
</tr>
<tr>
<td>World Wide Web (sites) [2]</td>
<td>153,127</td>
<td>35.2</td>
<td>0.11</td>
<td>0.00023</td>
</tr>
<tr>
<td>power grid [192]</td>
<td>4,941</td>
<td>2.7</td>
<td>0.080</td>
<td>0.00054</td>
</tr>
<tr>
<td>biology collaborations [140]</td>
<td>1,520,251</td>
<td>15.5</td>
<td>0.081</td>
<td>0.000010</td>
</tr>
<tr>
<td>mathematics collaborations [141]</td>
<td>253,339</td>
<td>3.9</td>
<td>0.15</td>
<td>0.000015</td>
</tr>
<tr>
<td>film actor collaborations [149]</td>
<td>449,913</td>
<td>113.4</td>
<td>0.20</td>
<td>0.00025</td>
</tr>
<tr>
<td>company directors [149]</td>
<td>7,673</td>
<td>14.4</td>
<td>0.59</td>
<td>0.0019</td>
</tr>
<tr>
<td>word co-occurrence [90]</td>
<td>460,902</td>
<td>70.1</td>
<td>0.44</td>
<td>0.00015</td>
</tr>
<tr>
<td>neural network [192]</td>
<td>282</td>
<td>14.0</td>
<td>0.28</td>
<td>0.049</td>
</tr>
<tr>
<td>metabolic network [69]</td>
<td>315</td>
<td>28.3</td>
<td>0.59</td>
<td>0.090</td>
</tr>
<tr>
<td>food web [138]</td>
<td>134</td>
<td>8.7</td>
<td>0.22</td>
<td>0.065</td>
</tr>
</tbody>
</table>
Graphs: paths

Path from node i to node j: a sequence of edges (directed or undirected from node i to node j)

- path length: number of edges on the path
- nodes i and j are connected
- cycle: a path that starts and ends at the same node
Graphs: shortest paths

Shortest Path from node i to node j

- also known as **BFS path**, or geodesic path
Measuring the small world phenomenon

d_{ij} = \text{shortest path between i and j}

Diameter:
\[ d = \max_{i,j} d_{ij} \]

Characteristic path length:
\[ \ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij} \]

Harmonic mean
\[ \ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1} \]
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C(1)$</th>
<th>$C(2)$</th>
<th>$r$</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.45</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20,416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>--</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105,323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>--</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107,182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,900</td>
<td>9.27</td>
<td>6.19</td>
<td>--</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311,313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>--</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311,313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td>--</td>
<td>0.005</td>
<td>0.001</td>
<td>--</td>
<td>8,9</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,000</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>136</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>--</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>--</td>
<td>0.005</td>
<td>0.001</td>
<td>--</td>
<td>0.029</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265,266</td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>2,695,904</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>--</td>
<td>0.067</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td></td>
<td></td>
<td></td>
<td>74</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td>3.0/--</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>351</td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5103</td>
<td>4.99</td>
<td>4.87</td>
<td></td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td>2.7</td>
<td></td>
<td>0.44</td>
<td>0.119</td>
<td>0.157</td>
<td>119,157</td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>--</td>
<td>0.189</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>--</td>
<td>0.10</td>
<td>0.080</td>
<td>--</td>
<td>0.003</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>--</td>
<td>0.69</td>
<td>0.033</td>
<td>--</td>
<td>0.366</td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>--</td>
<td>0.016</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>--</td>
<td>0.033</td>
<td>0.012</td>
<td>--</td>
<td>0.119</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,997</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>--</td>
<td>0.154</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>--</td>
<td>0.366</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>--</td>
<td>0.240</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>--</td>
<td>0.156</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td></td>
<td>0.16</td>
<td>0.23</td>
<td>--</td>
<td>0.263</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td></td>
<td>0.20</td>
<td>0.087</td>
<td>--</td>
<td>0.326</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td></td>
<td>0.18</td>
<td>0.28</td>
<td>--</td>
<td>0.226</td>
</tr>
</tbody>
</table>

**TABLE II** Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex-vertex distance $\ell$; exponent $\alpha$ of degree distribution if the distribution follows a power law (or $\alpha$ if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C(1)$ from Eq. (3); clustering coefficient $C(2)$ from Eq. (6); and degree correlation coefficient $r$, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Degree correlations

Do high degree nodes tend to link to high degree nodes?

Newman

compute the correlation coefficient of the degrees $x_i$, $y_i$ of the two endpoints of an edge $i$

$$r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \over \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td>–</td>
<td>0.16</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.0005</td>
<td>0.001</td>
<td>–</td>
<td>45</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>–</td>
<td>0.0005</td>
<td>0.001</td>
<td>–</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2810</td>
<td>8,9</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.16</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>WWW nd. edu</td>
<td>directed</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td>3.0/–</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td>2.7</td>
<td>–</td>
<td>0.44</td>
<td>0.119</td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>–0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>–</td>
<td>0.10</td>
<td>0.080</td>
<td>–0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.09</td>
<td>0.033</td>
<td>266</td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>2.1/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>–0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>–</td>
<td>0.033</td>
<td>0.012</td>
<td>–0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,997</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>–0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>–0.366</td>
<td>6, 354</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices $n$; total number of edges $m$; mean degree $z$; mean vertex–vertex distance $\ell$; exponent $\alpha$ of degree distribution if the distribution follows a power law (or “–” if not); in/out-degree exponents are given for directed graphs; clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient $r$, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Clique $K_n$

A graph that has all possible $n(n-1)/2$ edges
Directed Graph

Strongly connected graph: there exists a path from every i to every j

Weakly connected graph: If edges are made to be undirected the graph is connected
Web Structure: Connected Components

Study of Web Graph (Broder et al. 2000)

SCC = strongly connected component

..strongly connected core tends to have small diameter

Source: A.Z. Broder et al., WWW 2000
Summary: real network properties

Most nodes have only a small number of neighbors (degree), but there are some nodes with very high degree (power-law degree distribution)

- scale-free networks

If a node \( x \) is connected to \( y \) and \( z \), then \( y \) and \( z \) are likely to be connected

- high clustering coefficient

Most nodes are just a few edges away on average.

- small world networks
A “Canonical” Natural Network has...

*Few* connected components:
- often only 1 or a small number, indep. of network size

*Small* diameter:
- often a constant independent of network size (like 6)
- or perhaps growing only logarithmically with network size or even shrink?
- typically exclude infinite distances

A *high* degree of clustering:
- considerably more so than for a random network
- in tension with small diameter

A *heavy-tailed* degree distribution:
- a small but reliable number of high-degree vertices
- often of *power law* form

Question: Is it possible that there is a unifying underlying generative process?
Measuring and modelling network properties

Around 1999

- Watts and Strogatz, Dynamics and small-world phenomenon
- Faloutsos, On power-law relationships of the Internet Topology
- Kleinberg et al., The Web as a graph
- Barabasi and Albert, The emergence of scaling in real networks
Basics: the Erdős-Rényi Model: $G(N,p)$

All edges are *equally probable and appear independently*

NW size $N > 1$ and probability $p$: *distribution $G(N,p)$*

- each edge $(u,v)$ chosen to appear with probability $p$
- $N(N-1)/2$ trials of a biased coin flip

The usual *regime of interest* is when $p \sim 1/N$, $N$ is large

- e.g. $p = 1/2N$, $p = 1/N$, $p = 2/N$, $p = 10/N$, $p = \log(N)/N$, etc.
- in expectation, each vertex will have a “small” number of neighbors
- will then examine what happens when $N \to \infty$
- can thus study properties of *large networks* with *bounded degree*

*Degree distribution* of a typical $G$ drawn from $G(N,p)$:

- draw $G$ according to $G(N,p)$; look at a random vertex $u$ in $G$
- what is $\Pr[\deg(u) = k]$ for any fixed $k$?
- *Poisson distribution* with mean $l = p(N-1) \sim pN$
- Sharply concentrated; *not* heavy-tailed

Especially easy to *generate* NWs from $G(N,p)$
For any fixed $m \leq N(N-1)/2$, define distribution $G(N,m)$:

- choose *uniformly* at random from all graphs with *exactly* $m$ edges
- $G(N,m)$ is “like” $G(N,p)$ with $p = m/(N(N-1)/2) \sim 2m/N^2$
- this intuition can be made precise, and is correct
- if $m = cN$ then $p = 2c/(N-1) \sim 2c/N$
- mathematically trickier than $G(N,p)$
Evolution of a Random Network

We have a large number $n$ of vertices

We start randomly adding edges one at a time

At what time $t$ will the network:

- have at least one "large" connected component?
- have a single connected component?
- have "small" diameter?
- have a "large" clique?

How gradually or suddenly do these properties appear?
Random graph models $G(N,p)$ and $G(N,m)$: Recap

Model $G(N,p)$:
- select each of the possible edges independently with prob. $p$
- $\text{expected}$ total number of edges is $pN(N-1)/2$
- expected degree of a vertex is $p(N-1)$
- degree will obey a Poisson distribution ($\text{not}$ heavy-tailed)

Model $G(N,m)$:
- select $\text{exactly } m$ of the $N(N-1)/2$ edges to appear
- all sets of $m$ edges equally likely
Random graphs degree distributions

The $G(N,p)$ degree distribution follows a binomial

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

Assuming $z=np$ is fixed, as $n \to \infty$ $B(n,k,p)$ is approximated by a Poisson distribution

$$p(k) = P(k;z) = \frac{z^k}{k!} e^{-z}$$

Highly concentrated around the mean, with a tail that drops exponentially
Monotone Network Properties

Threshold or tipping for (say) connectivity:
- fewer than \( m(N) \) edges \( \rightarrow \) graph almost certainly \( \textit{not} \) connected
- more than \( m(N) \) edges \( \rightarrow \) graph almost certainly \( \textit{is} \) connected
- made formal by examining limit as \( N \rightarrow \infty \)

Often interested in \textit{monotone} graph properties:
- let \( G \) have the property
- \textit{add edges} to \( G \) to obtain \( G' \)
- then \( G' \) \textit{must} have the property also

Examples:
- \( G \) is connected
- \( G \) has diameter \( \leq d \) (\textit{not} exactly \( d \))
- \( G \) has a clique of size \( \geq k \) (\textit{not} exactly \( k \))
- \( d, k \) may depend on NW size \( N \) (How?)

Difficult to study emergence of non-monotone properties as the number of edges is increased - what would it mean?
Consider Erdos-Renyi $G(N,m)$ model

- select $m$ edges at random to include in $G$

Let $P$ be some *monotone* property of graphs

- $P(G) = 1 \rightarrow G$ has the property
- $P(G) = 0 \rightarrow G$ does not have the property

Let $m(N)$ be some function of network size $N$

- formalize idea that property $P$ appears “suddenly” at $m(N)$ edges

Say that $m(N)$ is a *threshold function for $P$* if:

- let $m'(N)$ be any function of $N$
- look at ratio $r(N) = m'(N)/m(N)$ as $N \rightarrow \infty$
- if $r(N) \rightarrow 0$: probability that $P(G) = 1$ in $G(N,m'(N))$: $\rightarrow 0$
- if $r(N) \rightarrow \infty$: probability that $P(G) = 1$ in $G(N,m'(N))$: $\rightarrow 1$

A *purely structural* definition of tipping

- tipping results from incremental increase in *connectivity*
Random Graphs: Which Properties Tip (2) ?

Connected component of size > N/2:
- threshold function is \( m(N) = \frac{N}{2} \) (or \( p \sim \frac{1}{N} \))
- note: full connectivity impossible

Fully connected:
- threshold function is \( m(N) = \frac{N/2}{\log(N)} \) (or \( p \sim \frac{\log(N)}{N} \))
- NW remains extremely sparse: only \( \sim \log(N) \) edges per vertex

Small diameter:
- threshold is \( m(N) \sim N^{3/2} \) for diameter 2 (or \( p \sim \frac{2}{\sqrt{N}} \))
- fraction of possible edges still \( \sim \frac{2}{\sqrt{N}} \rightarrow 0 \)
- generate very small worlds
Erdos-Renyi Summary

A model in which all connections are *equally likely*
- each of the $N(N-1)/2$ edges chosen randomly & independently

As we add edges, a *precise sequence* of events unfolds:
- graph acquires a giant component
- graph becomes connected
- graph acquires small diameter

Many properties appear *very suddenly* (tipping, thresholds)

All statements are *mathematically precise*

But is this how natural networks form?

If not, which aspects are unrealistic?
- may all edges are *not* equally likely!
Erdos-Renyi: Clustering Coefficient

Generate a network $G$ according to $G(N,p)$
Examine a “typical” vertex $u$ in $G$
  - choose $u$ at random among all vertices in $G$
  - what do we expect $c(u)$ to be?
Answer: exactly $p!$
In $G(N,m)$, expect $c(u)$ to be $\frac{2m}{N(N-1)}$
Both cases: $c(u)$ entirely determined by *overall* density
Baseline for comparison with “more clustered” models
  - Erdos-Renyi has *no bias* towards clustered or local edges
Watt’s models: Caveman and Solaria

Erdos-Renyi:

- sharing a common neighbor makes two vertices *no more likely* to be directly connected than two very “distant” vertices
- every edge appears entirely *independently* of existing structure

But in many settings, the *opposite* is true:

- you tend to meet new friends through your old friends
- two web pages pointing to a third might share a topic
- two companies selling goods to a third are in related industries

Watts’ *Caveman* world:

- *overall* density of edges is low
- but two vertices with a common neighbor are likely connected

Watts’ *Solaria* world

- overall density of edges low; no special bias towards local edges
- “like” Erdos-Renyi
Making it (Somewhat) Precise: the $\alpha$-model

The $\alpha$-model has the following parameters or “knobs”:

- $N$: size of the network to be generated
- $k$: the average degree of a vertex in the network to be generated
- $p$: the default probability two vertices are connected
- $\alpha$: adjustable parameter dictating bias towards local connections

For any vertices $u$ and $v$:

- define $m(u,v)$ to be the number of common neighbors (so far)

Key quantity: the propensity $R(u,v)$ of $u$ to connect to $v$

- if $m(u,v) \geq k$, $R(u,v) = 1$ (share too many friends not to connect)
- if $m(u,v) = 0$, $R(u,v) = p$ (no mutual friends $\rightarrow$ no bias to connect)
- else, $R(u,v) = p + (m(u,v)/k)^\alpha (1-p)$

Generate NW incrementally

- using $R(u,v)$ as the edge probability; details omitted

Note: $\alpha$ = infinity is “like” Erdos-Renyi (but not exactly)
Watts-Strogatz Model

$C(p) : \text{clustering coeff.}$

$L(p) : \text{average path length}$

(Watts and Strogatz, Nature 393, 440 (1998))
Small Worlds and Occam’s Razor

For small $\alpha$, should generate large clustering coefficients

- we “programmed” the model to do so
- Watts claims that proving precise statements is hard...

But we do not want a new model for every little property

- Erdos-Renyi $\rightarrow$ small diameter
- $\alpha$-model $\rightarrow$ high clustering coefficient

In the interests of Occam’s Razor, we would like to find

- a single, simple model of network generation...
- ... that simultaneously captures many properties

Watt’s small world: small diameter and high clustering
Meanwhile, Back in the Real World…

Watts examines three real networks as case studies:
- the Kevin Bacon graph
- the Western states power grid
- the nervous system

For each of these networks, he:
- computes its size, diameter, and clustering coefficient
- compares diameter and clustering to best Erdos-Renyi approximate
- shows that the best $\alpha$-model approximation is better
- important to be “fair” to each model by finding best fit

Overall moral:
- if we care only about diameter and clustering, $\alpha$ is better than $p$
Expected Result

\[ \langle k \rangle \sim 6 \]
\[ P(k=500) \sim 10^{-99} \]
\[ N_{\text{WWW}} \sim 10^9 \]
\[ \Rightarrow N(k=500) \sim 10^{-90} \]

Real Result

\[ P_{\text{out}}(k) \sim k^{-\gamma_{\text{out}}} \]
\[ P_{\text{in}}(k) \sim k^{-\gamma_{\text{in}}} \]

\[ \gamma_{\text{out}} = 2.45 \]
\[ \gamma_{\text{in}} = 2.1 \]

\[ P(k=500) \sim 10^{-6} \]
\[ N_{\text{WWW}} \sim 10^9 \]
\[ \Rightarrow N(k=500) \sim 10^3 \]

What does that mean?

Poisson distribution

Power-law distribution

Exponential Network

Scale-free Network
Scale-free Networks

The number of nodes (N) is not fixed
- Networks continuously expand by additional new nodes
  - WWW: addition of new nodes
  - Citation: publication of new papers

The attachment is not uniform
- A node is linked with higher probability to a node that already has a large number of links
  - WWW: new documents link to well known sites (CNN, Yahoo, Google)
  - Citation: Well cited papers are more likely to be cited again
Scale-Free Networks: the Barabasi-Albert model

**Growth:** Starting with a small number \((m_0)\) of nodes, at every time step, we add a new node with \(m(<< m_0)\) edges that link the new node to \(m\) different nodes already present in the system.

**Preferential attachment:** The probability \(\Pi\) that a new node will be connected to node \(i\) depends on the degree \(k_i\) of node \(i\), such that

\[
\Pi(k_i) = \frac{k_i}{\sum_i k_i}
\]

After \(t\) time-steps the network has \(N = t + m_0\) nodes and \(m \times t\) edges. Vertices \(j\) with high degree are likely to get more links ("Rich get richer")

The network evolves into a stationary scale-free state with

\[
P(k) \sim 2m^2 k^{-\gamma} \quad \gamma_{BA} = 3
\]
**Scale-Free Networks: Properties**

Preferential attachment explains

- heavy-tailed degree distributions
- small diameter ($\sim \log(N)$, via “hubs”)

Small average path length:  
$$\ell \sim \frac{\ln(N)}{\ln\ln(N)}$$

Node degree correlations: The dynamical process that creates a scale free network builds up nontrivial correlations between the degrees of connected nodes.

Clustering coefficient is ~5 times larger than that of a random graph. Will *not* generate a high clustering coefficient as with Watts-Strogatz model. Clustering coefficient decreases with network size, following approximately a power law  
$$C \sim N^{-0.75}$$

i.e. no bias towards local connectivity, but towards hubs.
Network robustness

The accidental failure of a number of nodes in a random network can fracture the system into non-communicating islands.

Scale-free networks are more robust in the face of such failures.

Scale-free networks are highly vulnerable to a coordinated attack against their hubs.