Models for Generating Social Networks

Prof. Shou-de Lin

CSIE/GINM, NTU

sdlin@csie.ntu.edu.tw
Administration Issues

• Homework (one each month):
  – You probably need to spend 20 hours or so on each homework.
  – HW 1.1 will be out today, 1.2 will be out next week. They will be due 3 weeks from now (10/13).
  – Graduate students: individual submission
  – Undergraduate students: can form a two-person team

• Final Project Group:
  – 3 people’s team is preferred
  – 2 persons’ team: please post on Ceiba to look for another member
  – 4 persons’ team: please break into 2*2 and follow the above rule
  – 1 person team: not allowed (come on, this is a SOCIAL network course. Don’t be so anti-social).
Reference Books for this Course

• Hanneman, Robert A. and Mark Riddle, Introduction to social network methods, (Online Text Book), Riverside, CA: University of California, Riverside, 2005.
Libraries for Network/Graph Programming

• JUNG (Java Universal Network/Graph Framework)
  – http://jung.sourceforge.net/
  – Rich intuitive sample codes in the examples

• The igraph library (C, R language)
  – http://igraph.sourceforge.net/
  – The most active and complete library using C for graphs
  – Rich examples and many graph theory algorithms

• The Boost Graph Library (C++)
  – Integrate generic programming standard template library

• LibSNA (Python)
  – http://www.libsna.org/
  – Provide simple data structure for graph manipulation

• Matlab
  – grTheory - Graph Theory Toolbox
    http://www.mathworks.com/matlabcentral/fileexchange/4266
  – Graph package
    http://www.mathworks.com/matlabcentral/fileexchange/12648
Review of last week: the property of natural social networks

1. There are generally two different definitions for network Diameter: (1) The largest shortest path. (2) The average length of all-pair shortest path (this lecture)
   – the average path lengths of many real-world social networks are small

2. High clustering coefficient (CC): the neighbors of a node are tightly connected.
   – To compute CC, we can simply use: \( \Sigma_i(\# \text{ of existing links between the neighbors } n_i / \# \text{ of possible links between neighbors of } n_i) \)

3. Degree distributions follow Power Law.

4. Contains patterns, motifs, groups, etc

The question to be answered in today’s class: How were social networks with the above characteristics formed? Can we model their generation process?
What is a Network Model?

• Informally, a network model is a **process** (randomized or deterministic) for generating a graph.

• Some people use “**generation model of networks**”.
Outline

• Random Graph
  – Erdos-Renyi Model
  – Configuration Model

• Scale-free Network
  – Power-law distribution
  – Barabasi-Albert Model

• Small-world Network
  – Watts-Strogatz Model

• Comparison of Network Models
Random Graph

For more details, please refer to:

Erdos-Renyi Random Graphs

(Or called ER model)

What does a “typical” graph with \( n \) vertices and \( m \) edges look like?

Paul Erdös (1913-1996)
Erdos-Renyi Random Graphs

• Consider a graph with \( n \) vertices
• Let \( E \) denote the total number of edge possible

\[
E = \binom{N}{2} = \frac{N!}{2!(N-2)!} = \frac{N(N-1)}{2}
\]

– (If directed, it would be multiply by 2)
Two Formulations of ER model

- \( G(n,p) \)
  - The ensemble of graphs constructed by putting in edges with probability \( p \), independent of one another (\( 1-p \) for absent edges)
  - Let \( G(n,p) \) be a random realization of \( G(n,p) \)
  - This is also called Poisson random graph

- \( G(n,m) \)
  - Randomly choose a graph from a set of all possible graphs with \( n \) nodes and \( m \) edges
  - Let \( G(n,m) \) be a random realization of \( G(n,m) \)

- Are these two models the same or different?

Note: the generation model produces a set of networks that follow certain distribution, however, it doesn’t mean in real implementation, you need to follow the same process to produce same type of networks.
Degree Distribution of $G(n,p)$

- Consider $G(n,p)$ for a **fixed** $p$ and $n$ (which is large)
- The absence or presence of an edge is **independent for all edges**
  - $\text{Prob}(\text{node } i \text{ connects to all other } n \text{ nodes}) = p^n$
  - $\text{Prob}(\text{node } i \text{ is isolated}) = (1-p)^n$
  - $\text{Prob}(\text{node } i \text{ has degree } k)$ follows a **binomial distribution**

$$p_k = \binom{n}{k} p^k (1 - p)^{n-k}$$
$G(n,p)$ is a Poisson random graph

- Define $z = np$ = average total degree

\[
\lim_{n \to \infty} p_k = \lim_{n \to \infty} \binom{n}{k} p^k (1 - p)^{n-k}
\]

\[
= \lim_{n \to \infty} \frac{n!}{(n-k)!k!} \left( \frac{z}{n} \right)^k \left( 1 - \frac{z}{n} \right)^{n-k}
\]

\[
= z^k e^{-z} / k!
\]

$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$

2009/9/29
Size of the Largest Component vs. p

- Let $C_{\text{max}}(p)$ be the number of nodes of the largest connected component of $G(n,p)$
- Apparently, $C_{\text{max}}(p)$ depends on $p$
  1. when $p$ is small, $C_{\text{max}}(p)$ is small $\Rightarrow$ average path length $= \text{infinite}$
  2. What happens when $p$ grows from 0 to 1?
For small $p$, few edges on the graph. Almost all vertices disconnected. The component size is small (no larger than $O(\log n)$, regardless of $p$).
Keep increasing $p$ ...
Keep increasing $p$ ...
Estimating the Size of the Largest Giant Component (LC)

- Let $u$ be the fraction of vertices that do NOT belong to LC
  - Choosing a vertex uniformly at random, $u$ is the probability that this vertex doesn’t belong to LC
- If a chosen vertex has degree $k$, then the probability that it does not belong to LC is equivalent to the probability that ”all its neighbors do not belong to LC”, which is $u^k$
- Since a chosen vertex can have degree from 0 to infinite, we can say
  \[ u = \sum_{k=0}^{\infty} p_k u^k = e^{-z} \sum_{k=0}^{\infty} \frac{z^k}{k!} u^k = e^{-z-u} \]
  Note that $e^x = \sum (x^k/k!)$
- The fraction $S$ of the graph occupied by a giant component is
  \[ S = 1 - u = 1 - e^{-zs} \]
- $S$ does not have an analytical solution (only numerical solution).
Evolution of a random graph

\begin{figure}
\centering
\includegraphics[width=\textwidth]{graph.png}
\caption{Evolution of a random graph for non-GCC vertices and giant component size.}
\end{figure}

mean component size $\langle s \rangle$

mean degree $z$

2009/9
P affects the Size of the Largest Component

- Erdos and Renyi proved that
  - $p < 1/n$: only **small** disconnected components
    - With probability tend to one as $N$ tens to infinity, the graph has no component of size greater than $O(\ln N)$, and no component has more than one cycle.
  - $p = 1/n$ (**phase transition**): almost surely the largest component has size $O(N^{2/3})$
  - $p > 1/n$: the graph has a component of $O(N)$, with a number $O(N)$ of cycles, and no other component has more than $O(\ln N)$ nodes and more than one cycle
  - When $p > \ln N/N$, then almost surely the graph is fully connected

Emergence of the Giant Component

- At \( z=1 \) (or \( pn = 1 \))
  - **Suddenly** the largest component contains a finite fraction \( F \) of the total number of vertices, \( C_{max} = FN \)
Giant Component in the Real World

- Many real-world networks “gain critical mass”
- E.g. WWW

Average Path Length of $G(n,p)$ when $p \gt \ln N/N$

- **When** $p > \ln N/N$, the graph is almost totally connected, when the average # of neighbors is $z = (n - 1)p \approx np$
- **Assuming the average path length is** $l$, then $z^l \approx n$,

$$l \approx \log(n) / \log(z)$$

In many real-world **small-world** social networks, $l \approx O(\log(n))$
The Clustering Coefficient for G(n,p)

- CC stands for how likely the two neighbors of a node is connected to each other.
- CC=p for G(n,p) since whether the two neighbors are connected is independent of whether they belong to the neighbors of a node.
- p goes to zero when n becomes large and z=1
## Clustering Coefficient for Different Networks

<table>
<thead>
<tr>
<th>Network</th>
<th>$n$</th>
<th>$z$</th>
<th>Measured Clustering Coefficient $C$</th>
<th>Random Graph Clustering Coefficient $C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet (autonomous systems)$^a$</td>
<td>6374</td>
<td>3.8</td>
<td>0.24</td>
<td>0.00060</td>
</tr>
<tr>
<td>World-Wide Web (sites)$^b$</td>
<td>153,127</td>
<td>35.2</td>
<td>0.11</td>
<td>0.00023</td>
</tr>
<tr>
<td>Power grid$^c$</td>
<td>4,941</td>
<td>2.7</td>
<td>0.080</td>
<td>0.00054</td>
</tr>
<tr>
<td>Biology collaborations$^d$</td>
<td>1,520,251</td>
<td>15.5</td>
<td>0.081</td>
<td>0.000010</td>
</tr>
<tr>
<td>Mathematics collaborations$^e$</td>
<td>253,339</td>
<td>3.9</td>
<td>0.15</td>
<td>0.000015</td>
</tr>
<tr>
<td>Film actor collaborations$^f$</td>
<td>449,913</td>
<td>113.4</td>
<td>0.20</td>
<td>0.00025</td>
</tr>
<tr>
<td>Company directors$^f$</td>
<td>7,673</td>
<td>14.4</td>
<td>0.59</td>
<td>0.0019</td>
</tr>
<tr>
<td>Word co-occurrence$^g$</td>
<td>460,902</td>
<td>70.1</td>
<td>0.44</td>
<td>0.00015</td>
</tr>
<tr>
<td>Neural network$^c$</td>
<td>282</td>
<td>14.0</td>
<td>0.28</td>
<td>0.049</td>
</tr>
<tr>
<td>Metabolic network$^h$</td>
<td>315</td>
<td>28.3</td>
<td>0.59</td>
<td>0.090</td>
</tr>
<tr>
<td>Food web$^i$</td>
<td>134</td>
<td>8.7</td>
<td>0.22</td>
<td>0.065</td>
</tr>
</tbody>
</table>
Properties of E-R Random Graphs

1. **Phase transition** in connectivity at average node degree, $z = 1$ (i.e., $p=1/n$)

2. Poisson degree distribution, $p_k = z^k e^{-z}/k!$

3. Diameter, $d \sim \log n$

4. **Clustering coefficient**: none ($C = p$)
Can $G(n,p)$ Model Real-world Networks?

1. **Phase transition**: **YES**!
   - The emergence of a giant component

2. **Poisson degree distribution**: **NO**!
   - Most networks are power-law distribution

3. **Small-world diameter**: **YES**!
   - It occurs in social, technological, knowledge, and biological networks

4. **High Clustering coefficient**: **NO**!
Outline

• Random Graph
  – Erdos-Renyi Model
  – Configuration Model

• Scale-free Network
  – Power-law distribution
  – Barabasi-Albert Model

• Small-world Network
  – Watts-Strogatz Model

• Comparison of Network Models
Configuration Model

More details please see
Configuration Model

• In configuration model, we are not explicitly specify how many links are there as $G(n,m)$ nor the fixed probability as $G(n,p)$

• Instead, we specify a degree distribution $p_k$, such that $p_k$ is the fraction of vertices in the network having degree $k$.

• How to generate a network that satisfies this distribution?
Phrase Transition in a Configuration Model

- Define $q_k$ as the **excess degree distribution**, which is the probability that in the end of a randomly picked edge connects a node with degree $k + 1$. $q_k = (k+1)p_{k+1}/\sum xp_x$

- The average number of vertices two steps away from a given vertex is $\sum kq_k$, which is
  \[ \sum_{k=0}^{\infty} kq_k = \sum_{k=0}^{\infty} \frac{k(k+1)p_{k+1}}{\sum x p_x} = \frac{\sum (k-1)kp_k}{\sum x p_x} = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = \frac{z_2}{z_1} \]

- Multiplying by the mean degree $z_1 = \langle k \rangle$, we can obtain the mean number of second neighbors of a vertex is $z_2 = \langle k^2 \rangle - \langle k \rangle$.

- Note that $z_2$ for Poisson random graph is $\langle k^2 \rangle$.

- Similarly, the mean $m$-step neighbors
  \[ z_m = (z_2/z_1)^m z_{m-1} = [z_2/z_1]^{m-1} z_1 \]

- If $z_2 > z_1$, then the graph will diverge exponentially.

- The boundary for phrase transition is $z_2 - z_1$, or equivalently $\sum_k k(k-2)p_k$
  - if larger than zero than there exists a large component
Component Size for Configuration Models

• Assuming $z_2 > z_1$, then the giant component $S$ exists.

• Since the mean m-step neighbors $z_m = [z_2/z_1]^{m-1} z_1$ grows exponentially with m, therefore we can roughly assume the average path length $\ell^m = n = z_m$ 
  $\Rightarrow \ell = \ln(n/z_1)/\ln(z_2/z_1) + 1$

• Recall for ER model, $\ell = \ln(n)/\ln(z)$

• The mean component size below the phrase transition in the region where there is no giant component is 
  $1 + z_1(1 + (z_2/z_1) + (z_2/z_1)^2 + ...) = 1 + z_1^2/(z_1 - z_2)$

• Define $u$ to be the probability that a randomly chosen edge leads to a vertex that is not part of the giant component.

• $u = \sum_{k=0}^{\infty} q_k u^k \quad S = 1 - \sum_{k=0}^{\infty} p_k u^k$
Generating the Configuration Model
(Bender-Canfield model, 1978)

• Given: a degree sequence \([d_1, d_2, \ldots, d_n]\)

• Algorithm
  1. Create \(d_i\) copies of node \(i\)
     • Each degree’s node set is called “stubs” or “spokes”
  2. Take a random pairing of the copies
     • With equal probability
     • Allow self-loops and multiple edges

• Output: the ensemble of graphs so produced

Note: this method is not necessary the most efficient one
Configuration Model (cont.)

- Suppose the degree sequence [4, 3, 2, 1]

1. Create **multiple copies** of the nodes according to degree

2. Pair the nodes uniformly at random

- Generate the resulting graph
Clustering Coefficient for Configuration Model

- In configuration model, the CC highly depends on the degree distribution → therefore we cannot say too much about it without knowing the distribution.
- We can try to calculate the average CC for Bender-Canfield method.
- The “average degree minus 1” of the neighbor nodes is \( \Sigma kq_k \).
- CC: probability one neighbor node is connecting to another neighbor node is roughly
  \[
  \left( \Sigma kq_k \right)^2 / nz_1 = \left( \frac{z_2}{z_1} \right)^2 / nz_1 = \left( \frac{z}{n} \right) * \left( \frac{z_2}{z_1} \right)^2
  \]
- That says, if \( n \) is very large, CC tends to be zero 😞 (but if \( \frac{z_2}{z_1} \) is large, then C can be non-negligible.)
Outline

• Random Graph
  – Erdos-Renyi Model
  – Configuration Model

• Scale-free Network
  – Power-law distribution
  – Barabasi-Albert Model

• Small-world Network
  – Watts-Strogatz Model

• Comparison of Network Models
Power Law Distribution

For more details, please refer to:

Normal (i.e., Gaussian) distribution of human heights

Average value close to most typical

Distribution close to symmetric around average value
Power-law Distribution

- High skew (asymmetry)
- Nearly straight line on a log-log plot
Heavy Tailed Distribution?

• Right skew
  – **Normal distribution** (not heavy tailed)
    • e.g., human height
  – **Zipf’s or power-law distribution** (heavy tailed)
    • e.g., city population sizes: very few cities with large size (e.g. NYC 8 million), but many small towns (e.g. size≈10^4)

• High ratio of max to min
  – E.g., Human height
    • 272cm to 57cm, ratio=4.8
  – E.g., City size
    • NYC 8 million to smallest town 52, ratio=150,000
Power Law is Ubiquitous!

Moby Dick (白鯨記) scientific papers 1981-1997 AOL users visiting sites 1997

(word frequency) (citations) (web hits)

10^0 10^2 10^4

10^4

10^2

10^0

bestsellers 1895-1965 AT&T customers on 1 day California 1910-1992

books sold telephone calls received earthquake magnitude

10^6 10^1 10^7

10^0 10^2 10^4 10^6

10^3 10^2

10^0

2 3 4 5 6 7

2009/9/29 SNA09, Modeling, Prof. Sd Lin
More Power laws...

Moon

Solar flares

wars (1816-1980)

(g)

(h)

(i)

\[ \text{crater diameter in km} \]

\[ \text{peak intensity} \]

\[ \text{intensity} \]

\[ \text{net worth in US dollars} \]

\[ \text{name frequency} \]

\[ \text{population of city} \]

richest individuals 2003

US family names 1990

US cities 2003

\(2009/9/29\)
Power Law Distribution

\[ p(x) = C x^{-\alpha} \]

Normalization constant

(probabilities over all \( x \) must sum to 1)

- Straight line on a log-log plot

\[ \ln(p(x)) = c - \alpha \ln(x) \]

- powers of a number will be uniformly spaced

\[ 2^0=1, 2^1=2, 2^2=4, 2^3=8, 2^4=16, 2^5=32, 2^6=64, \ldots \]
Fitting Power-law Distribution

• Most common and not very accurate method
  – Bin the different values of $x$ and create a frequency histogram

\[ \ln(\text{# of times } x \text{ occurred}) \]

\[ x \text{ stands for various quantities} \]

$\ln(x)$ is the natural logarithm of $x$, but any other base of the logarithm will give the same exponent of a since

\[ \log_{10}(x) = \frac{\ln(x)}{\ln(10)} \]
Log-log Scale Plot of Straight Binning

• Same bins, but plotted on a log-log scale

- Tens of thousands of observations for $x < 10$

- Noise in the tail
  Only 0, 1 or 2 observations for $x > 500$
Log-log Scale Plot of Straight Binning (cont.)

- Fitting a straight line to it via least squares regression will overfit the low value ones.

Few bins vs. much more bins.
Solution-1: Logarithmic Binning

- **Normalization**
  - Bin data into *exponentially wider bins*
  - $1, 2, 4, 8, 16, 32, \ldots$

Disadvantage: lose information
Solution-2: Cumulative Binning

• No loss of information
• Take advantage of cumulative distribution
  – How many observations are at least $x$? $P_{cd}(x) = \int_{x}^{\infty} x^{-\alpha} \, dx$
  – The cumulative probability of a power-law distribution is also power-law but with an exponent $\alpha - 1$

$$P_{cd}(x) = \int_{x}^{\infty} x^{-\alpha} \, dx = \frac{c}{1 - \alpha} x^{-(\alpha - 1)}$$
Fitting via Regression to the Cumulative Distribution

$\alpha - 1 = 1.43$ fit
Where to Start Fitting?

- Some data exhibit a power law only in the tail
  - Binning and cumulative distribution only fit tails
- Select an $x_{min}$ where the power-law could start
  - E.g., distribution of paper citations: $x_{min} > 100$
Normalize C of Power Law

\[ p(x) = Cx^{-\alpha} \quad (\alpha > 0) \]

Probabilities over all \( x \) must sum to 1

- There must be some lowest value \( x_{\text{min}} \) at which the power law is obeyed. We start there.

\[
1 = \int_{x_{\text{min}}}^{\infty} p(x) \, dx = C \int_{x_{\text{min}}}^{\infty} x^{-\alpha} \, dx = \frac{C}{1 - \alpha} \left[ x^{-\alpha+1} \right]_{x_{\text{min}}}^{\infty}
\]

\[
C = (\alpha - 1)x_{\text{min}}^{\alpha-1} \quad (\alpha > 1)
\]

\[
p(x) = \frac{\alpha - 1}{x_{\text{min}}} \left( \frac{x}{x_{\text{min}}} \right)^{-\alpha}
\]
Solution-3: Maximum Likelihood Fitting

\[ p(x) = C x^{-\alpha} = \frac{\alpha - 1}{x_{\text{min}}} \left( \frac{x}{x_{\text{min}}} \right)^{-\alpha} \]

- Given a dataset containing \( n \) observations \( x_i \geq x_{\text{min}} \), use the **maximum likelihood estimator** (MLE) to find \( \alpha \) for the power-law model that is most likely to have generated the given data

\[ P(x|\alpha) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\text{min}}} \left( \frac{x_i}{x_{\text{min}}} \right)^{-\alpha} \]

- The **likelihood** of the data given the model
  - **Maximize** this function!!
Maximum Likelihood Fitting (cont.)

- Commonly the logarithmic $L$ of the likelihood is used for parameter estimation

$$
L = \ln p(x | \alpha) = \ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\text{min}}} \left( \frac{x_i}{x_{\text{min}}} \right)^{-\alpha}
$$

$$
= \sum_{i=1}^{n} \left[ \ln(\alpha - 1) - \ln x_{\text{min}} - \alpha \ln \frac{x_i}{x_{\text{min}}} \right]
$$

$$
= n \ln(\alpha - 1) - n \ln x_{\text{min}} - \alpha \sum_{i=1}^{n} \ln \frac{x_i}{x_{\text{min}}}
$$

Setting $\partial L / \partial \alpha = 0$ and solving for $\alpha$:

$$
\hat{\alpha} - 1 + n \left[ \sum_{i=1}^{n} \ln \frac{x_i}{x_{\text{min}}} \right]^{-1}
$$
## Some Exponents for Real-world Data [Newman’05]

<table>
<thead>
<tr>
<th>quantity</th>
<th>minimum $x_{\text{min}}$</th>
<th>exponent $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) frequency of use of words</td>
<td>1</td>
<td>2.20</td>
</tr>
<tr>
<td>(b) number of citations to papers</td>
<td>100</td>
<td>3.04</td>
</tr>
<tr>
<td>(c) number of hits on web sites</td>
<td>1</td>
<td>2.40</td>
</tr>
<tr>
<td>(d) copies of books sold in the US</td>
<td>2,000,000</td>
<td>3.51</td>
</tr>
<tr>
<td>(e) telephone calls received</td>
<td>10</td>
<td>2.22</td>
</tr>
<tr>
<td>(f) magnitude of earthquakes</td>
<td>3.8</td>
<td>3.04</td>
</tr>
<tr>
<td>(g) diameter of moon craters</td>
<td>0.01</td>
<td>3.14</td>
</tr>
<tr>
<td>(h) intensity of solar flares</td>
<td>200</td>
<td>1.83</td>
</tr>
<tr>
<td>(i) intensity of wars</td>
<td>3</td>
<td>1.80</td>
</tr>
<tr>
<td>(j) net worth of Americans</td>
<td>$600m$</td>
<td>2.09</td>
</tr>
<tr>
<td>(k) frequency of family names</td>
<td>10,000</td>
<td>1.94</td>
</tr>
<tr>
<td>(l) population of US cities</td>
<td>40,000</td>
<td>2.30</td>
</tr>
</tbody>
</table>
### Exponents for Real-world Network [Newman’03]

<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>ℓ</th>
<th>α</th>
<th>(C^{(1)})</th>
<th>(C^{(2)})</th>
<th>r</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>social</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449913</td>
<td>25516482</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>55392</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>253339</td>
<td>496480</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>52909</td>
<td>245300</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>1520251</td>
<td>11803064</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>47000000</td>
<td>80000000</td>
<td>3.16</td>
<td>–</td>
<td>2.1</td>
<td>0.16</td>
<td>0.15</td>
<td>–</td>
<td>8, 9</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59912</td>
<td>86300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16881</td>
<td>57029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>–</td>
<td>45</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>–</td>
<td>0.029</td>
<td>–</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2810</td>
<td>–</td>
<td></td>
<td></td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>information</td>
<td></td>
<td></td>
<td></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>269504</td>
<td>1497135</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>203549046</td>
<td>2130000000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.74</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783339</td>
<td>6716198</td>
<td>8.57</td>
<td>3.0/–</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>351</td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1022</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>460902</td>
<td>17000000</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>technological</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10607</td>
<td>31992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>–0.180</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4941</td>
<td>6594</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>19603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.69</td>
<td>–0.033</td>
<td>–0.033</td>
<td>366</td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1439</td>
<td>1723</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>1377</td>
<td>2213</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>24097</td>
<td>53248</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>1296</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>biological</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.000</td>
<td>0.67</td>
<td>–0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2115</td>
<td>2240</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>2399</td>
<td>7.66</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>
80/20 Rule

- The fraction $W$ of the wealth in the hands of the richest $P$ of the population is given by
  \[ W = P^{(\alpha-2)/(\alpha-1)} \]
  - E.g., US wealth $\alpha = 2.1$

- The richest 20% holds 86% of the wealth
Distribution that do **NOT** follow power law

- **Ex-(a):** The abundance of North American bird species
  - Though it spans over five order of magnitude
  - It follows a **log-normally distribution**
    \[ f_X(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0 \]

- **Ex-(b):** The number of entries in people’s email address books
  - Though it spans about three orders of magnitude
  - It follows a **stretched exponential**
    \[ p(x) = e^{-ax^b} \]

- **Ex-(c):** The distribution of the sizes of forest fires
  - Though it spans six orders of magnitude
  - It follows a **power law with an exponential cutoff**
    \[ p(x) = Cx^{-\alpha}e^{-\lambda x} \]
Outline

• Random Graph
  – Erdos-Renyi Model
  – Configuration Model

• Scale-free Network
  – Power-law distribution
  – Barabasi-Albert Model

• Small-world Network
  – Watts-Strogatz Model

• Comparison of Network Models
Scale-free Network & Barabasi-Albert Model

For more details, please refer to:

Scale-Free Network

- A network whose **degree distribution** follows a **power law**, at least asymptotically (漸進線地)

\[ P(k) = Ck^{-\alpha} \implies P(k) \sim k^{-\alpha} \]

- Typically \(2 < \alpha < 3\), occasionally lie outside bounds

- Q: How to model or generate the scale-free networks?
Poisson vs. Scale-free network

Poisson network
(Erdos-Renyi random graph)

Degree distribution is Poisson

Scale-free network
(power-law network)

Degree distribution is Power-law

2009/9/29 SNA09, Modeling, Prof. Sd Lin
Should the number of nodes be fixed?

• Random graph models such as ER assume the fixed number of nodes, and the probability that two nodes are connected is independent of the nodes’ degree.

• However, most real-world networks grow by continuously adding new nodes
  – E.g., WWW, citations

• Can we model such network growing?
Model for Exponential Network

• Assuming initially, there are two connected nodes. Then with each increment of time, a new vertex is added into the network and connects to a randomly chosen old vertex (without preference).
  – Therefore at time t, there shall be t+1 vertices and t edges. The total degree is 2t
• Assume the probability $p(k,s,t)$ stands for the probability that a vertex $s$ has degree $k$ at time $t$. Then
  $$p(k,s,t+1) = \frac{1}{t+1} p(k-1,s,t) + \left(1 - \frac{1}{t+1}\right) p(k,s,t) \quad \text{...eq1}$$
• The total degree distribution of the entire network is
  $$p(k,t) = \sum_s p(k,s,t)/(t+1)$$
• Apply $\sum_s$ to both sides of eq1, we get
  $$(t+2)p(k,t+1) = p(k-1,t) + tp(k,t)$$
• When $t \to \infty$, the above equation becomes $2p(k,t) = p(k-1,t)$, which implies $p(k,t)$ is of the form $2^{-k} \quad \Rightarrow \quad \text{Exponential Network}$
• Unfortunately, real-world networks follow power law.
Barabasi-Albert Model (1/2) [1999]

• Idea: Real-world networks follow preferential attachment
  – New nodes tend to connect themselves to a given node that has higher degree.
  – A newly created website is more likely to be linked to a website that is linked by many other people — popularity is attractive.

• Following such idea, Barabasi & Albert proposes $p_{m,t}$ as the probability an existing node $m$ is connected to a newly added node at time $t$, and set $p_{m,t}=k/2t$, where $k$ is the degree of $m$ at time $t$. 
Barabasi-Albert Model (2/2)\textsuperscript{[1999]}

- Similar to exponential network model
  \[ p(k,s,t+1) = \frac{k-1}{2t} p(k-1,s,t) + \left(1 - \frac{k}{2t}\right) p(k,s,t) \ldots eq2 \]

- Similar to the exponential network model, since
  \[ p(k,t) = \sum_s p(k,s,t)/(t+1) \]
  we can sum over \( s \) on both sides of eq2, and let \( t \to \infty \), to obtain
  \[ p(k,t)+1/2[k*p(k,t)-(k-1)*p(k-1,t)]=0 \]

- The solution is of the form \( k^{-3} \to \) power law
A more General Barabasi-Albert Model

1. **Start** with a small number \((m_0)\) of fully connected nodes

2. Add one new node with \(m (\leq m_0)\) edges
   - Link the new node to \(m\) different nodes already present in the system
   - Based on the degree \(k_i\) of node \(I\)

3. After \(t\) time steps
   - There are \(N=t+m_0\) nodes,
     - \(mt + C_{m_0}^2\) edges
     - total degree = \(2mt + 2C_{m_0}^2\)
Generating Barabasi-Albert Network

• Start with a set of $m_0=3$ fully connected nodes

• Add a new node 4, it has $m=2$ edges
  – Prob(selecting any node)=1/3
  – Suppose they are node 2 and node 3

• Add a new node 5, it has $m=2$ edges
  – Prob(selecting node 1)=2/10=1/5
  – Prob(selecting node 2)=3/10
  – Prob(selecting node 3)=3/10
  – Prob/selecting node 4)=2/10=1/5

• Add a new node 6...
  – 2/14, 3/14, 4/14, 3/14, 2/14
Continuum Theory (1/4)

- Proving the general Barabasi-Albert model follows $k^{-3}$
- Assume the degree distribution $k_i$ is a continuous variable
  - The rate at which $k_i$ change is expected to be proportional to
    \[
    \Pi(k_i) = \frac{k_i}{\sum_j k_j}
    \]
    \[
    \frac{\partial k_i}{\partial t} = m\Pi(k_i) = m \frac{k_i}{2m(t-1)+C_2^{m_0}} \approx \frac{k_i}{2t}
    \]
Continuum Theory (2/4)

\[ \frac{\partial k_i}{\partial t} = \frac{k_i}{2t} \]

• Solve with Initial condition each node i at its introduction has \( k_i(t_i) = m \)

\[ k_i(t) = m \left( \frac{t}{t_i} \right)^{0.5} \]

The later a node is added, the smaller its degree is.

• The probability a node has a degree \( k_i(t) \) smaller than \( k \) is

\[ P(k_i(t) < k) = P \left( t_i > \frac{m^2 t}{k^2} \right) \]
Continuum Theory (3/4)

\[ P(k_i(t) < k) = P \left( t_i > \frac{m^2 t}{k^2} \right) \]

- Assume we add the nodes at equal time intervals, \( t_i \) has a constant probability density

\[ P(t_i) = \frac{1}{m_0 + t} \]

\[ P(k_i(t) < k) = P \left( t_i > \frac{m^2 t}{k^2} \right) \]

\[ = 1 - P \left( t_i \leq \frac{m^2 t}{k^2} \right) = 1 - \frac{m^2 t}{k^2(t + m_0)} \]
Theoretical Approach

Continuum Theory (4/4)

• Finally, the degree distribution $P(k)$ can be obtained using

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{2m^2t}{m_0 + t} \frac{1}{k^3}$$

$$P(k) \sim 2m^2k^{-3} \quad \text{c.f. } P(k) \sim k^{-\alpha}$$

Independent of $m$, $N=m_0 + t$, i.e., independent of continuous growth in agreement with the numerical simulations

\[ \alpha = 3 \]

\[ \Rightarrow \quad \text{Time-independent degree distribution} \]
See $P(k)$

Fix: $N$, Vary: $m_0$, $m$

- $N=m_0+t=300,000$
  - $m_0=m=1$  
  - $m_0=m=3$  
  - $m_0=m=5$  
  - $m_0=m=7$
- Slope (dashed line) = 2.9  
  - i.e., $\alpha=2.9$
See $P(k)$

Vary: $N$, Fix: $m_0$, $m$

- $m_0=m=5$
  - $N=100,000$
  - $N=150,000$
  - $N=200,000$
- Slope (dashed line) = 2.9
  - i.e., $\alpha = 2.9$
Average Path length of Scale-free Networks

• **Average Path length** \( (L) \)
  - If \( \alpha > 3 \) \( \Rightarrow L \approx O(\log(N)/\log(z_2/z_1)) \)
  - If \( 2 < \alpha < 3 \) \( \Rightarrow L \approx O(\log\log N) \), given the average degree is strictly greater than 1 and the maximum degree is sufficiently large.

Clustering Coefficient of Scale-free Networks

- Using $CC = \frac{(z/n) \cdot (z_2/z_1^2)^2}{z_2^2}$, it is possible to obtain $CC \approx N^{(7-3\alpha)/(\alpha-1)}$
  - If $\alpha>7/3$, then $CC$ tends to become 0 with large $N$
  - If $\alpha=7/3$, then $CC =1$
  - If $\alpha<7/3$, then $CC$ grows with $N$

Overview: Properties of BA-Model

• Power-law degree distribution
• The network is connected
  – Every node is born with a link (m=1) or several links (m>1)
• Small network diameter (at most \(O(\log N)\))
• Usually small clustering coefficient (large enough \(\alpha\))
• The older get richer
  – Preferential attachment
  – Nodes accumulate links as time goes on
Outline

• Random Graph
  – Erdos-Renyi Model
  – Configuration Model

• Scale-free Network
  – Power-law distribution
  – Barabasi-Albert Model

• Small-world Network
  – Watts-Strogatz Model

• Comparison of Network Models
Small-world Network

For more details, please refer to:

Small-world Phenomenon

- “Six Degree of Separation”
- **Milgram’s Experiment** [1969]
  - 64/296 (25%) reach
  - Average path length = 6
## Observation in Real-world Cases (1/2)

<table>
<thead>
<tr>
<th>Network</th>
<th>Size</th>
<th>$\langle k \rangle$</th>
<th>$\ell \sim \ell_{\text{rand}}$</th>
<th>$C &gt; C_{\text{rand}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW, site level, undir.</td>
<td>153,127</td>
<td>35.21</td>
<td>3.1 - 3.35</td>
<td>0.1078 0.00023</td>
</tr>
<tr>
<td>Internet, domain level</td>
<td>3015 - 6209</td>
<td>3.52 - 4.11</td>
<td>3.7 - 3.76 6.36 - 6.18</td>
<td>0.18 - 0.3 0.001</td>
</tr>
<tr>
<td>Movie actors</td>
<td>225, 226</td>
<td>61</td>
<td>3.65 2.99</td>
<td>0.79 0.00027</td>
</tr>
<tr>
<td>LANL coauthorship</td>
<td>52, 909</td>
<td>9.7</td>
<td>5.9 4.79</td>
<td>0.43 1.8 × 10^{-4}</td>
</tr>
<tr>
<td>MEDLINE coauthorship</td>
<td>1,520, 251</td>
<td>18.1</td>
<td>4.6 4.91</td>
<td>0.066 1.1 × 10^{-5}</td>
</tr>
<tr>
<td>SPIRES coauthorship</td>
<td>56, 627</td>
<td>173</td>
<td>4.0 2.12</td>
<td>0.726 0.003</td>
</tr>
<tr>
<td>NCSTRL coauthorship</td>
<td>11,994</td>
<td>3.59</td>
<td>9.7 7.34</td>
<td>0.496 3 × 10^{-4}</td>
</tr>
<tr>
<td>Math coauthorship</td>
<td>70, 975</td>
<td>3.9</td>
<td>9.5 8.2</td>
<td>0.59 5.4 × 10^{-5}</td>
</tr>
<tr>
<td>Neurosci. coauthorship</td>
<td>209, 293</td>
<td>11.5</td>
<td>6 5.01</td>
<td>0.76 5.5 × 10^{-5}</td>
</tr>
<tr>
<td><em>E. coli</em>, substrate graph</td>
<td>282</td>
<td>7.35</td>
<td>2.9 3.04</td>
<td>0.32 0.026</td>
</tr>
<tr>
<td><em>E. coli</em>, reaction graph</td>
<td>315</td>
<td>28.3</td>
<td>2.62 1.98</td>
<td>0.59 0.09</td>
</tr>
<tr>
<td>Ythan estuary food web</td>
<td>134</td>
<td>8.7</td>
<td>2.43 2.26</td>
<td>0.22 0.06</td>
</tr>
<tr>
<td>Silwood park food web</td>
<td>154</td>
<td>4.75</td>
<td>3.40 3.23</td>
<td>0.15 0.03</td>
</tr>
<tr>
<td>Words, cooccurrence</td>
<td>460,902</td>
<td>70.13</td>
<td>2.67 3.03</td>
<td>0.437 0.0001</td>
</tr>
<tr>
<td>Words, synonyms</td>
<td>22,311</td>
<td>13.48</td>
<td>4.5 3.84</td>
<td>0.7 0.0006</td>
</tr>
<tr>
<td>Power grid</td>
<td>4,941</td>
<td>2.67</td>
<td>18.7 12.4</td>
<td>0.08 0.005</td>
</tr>
<tr>
<td><em>C. Elegans</em></td>
<td>282</td>
<td>14</td>
<td>2.65 2.25</td>
<td>0.28 0.05</td>
</tr>
</tbody>
</table>

2009/9/29  SNA09, Modelling  Prof. Sd Lin
Observation in Real-world Cases (2/2)

\[ \ell_{\text{rand}} \sim \frac{\ln(N)}{\ln(\langle k \rangle)} \]

\[ C_{\text{rand}} = p = \frac{\langle k \rangle}{N} \]

- **Average Path Length**
  \[ \ell_{\text{rand}} \sim \ell_{\text{real-world}} \]

- **Clustering Coefficient**
  \[ C_{\text{real-world}} \text{ appears to be independent of the network size} \]

Characteristic of **Ring Lattice**!
• Ring lattice with \( N \) nodes
  – Each has \( k \) neighbors
  – \(|E| = kN/2\)

• Longest Path Length: \( N/K \)
• Average Path Length (\( l \)): \( N/2K \)

• Clustering Coefficient

\[
C = \frac{C_2^k - (1 + 2 + ... k/2)}{C_2^k} = \frac{3(k-2)}{4(k-1)} \approx \frac{3}{4}
\]

Increase much faster than random and real-world networks

Independent of network size

2009/9/29
Watts-Strogatz Model [1998]

Summary of observations

• One extreme: **Random graph**
  – Low avg path length, **low** clustering coefficient

• Other extreme: “**Regular**” network (**Lattice**)
  – **High** avg path length, **high** clustering coefficient

• Real-world case: **Small-world Network**
  – **Low** avg path length, **high** clustering coefficient
Watts-Strogatz Model (cont.)

• **Interpolate** between lattice and random graph!

1. **Start with order**
   - Start with a ring lattice with \( N \) nodes (\( K \) neighbors)
   - To ensure sparse and connected network
     - \( N \gg K \gg \ln(N) \gg 1 \)

2. **Random Rewiring** → **shortcuts**
   - Randomly rewire each edge with probability \( p \)
   - Disallow self-loop and duplicate edges
   - Introduce \( pNK/2 \) long-range edges
Watts-Strogatz Model (cont.)

Regular

Small-world

Random

$p = 0$  
Increasing randomness  
$p = 1$
Basic Idea of WS Model

1. Most people are friends with their immediate neighbors (**neighbored edges**) 
   – Neighbors on the same street, colleagues, etc.

2. Everyone has one or two friends who are far away (**long-range edges by rewiring**) 
   – People in other countries, old acquaintances 
   – Such tie plays the role of “short cut”, and is critical to small world phenomenon.
Characteristic $l(p)$ and $C(p)$ for WS Model

Coexistence of small $l$ and large $C$

$\begin{align*}
l(0) &\sim N/2K \\
C(0) &\sim 3/4
\end{align*}$

$\begin{align*}
l(1) &\sim \log(N)/\log(K) \\
C(1) &\sim K/N
\end{align*}$
Average Path Length for WS Model

- \( L = \frac{N}{2K} \) when \( p=0 \), and \( L \) does not begin to decrease significantly until \( p \geq \frac{2}{NK} \), which implies more than one shortcut occurs.

- The above implies the transition \( p \) depends on \( N \). That is, there exist a crossover size \( N^* \) such that if \( N < N^* \), \( L \propto N \); and if \( N > N^* \), \( L \propto \ln(N) \)

- It has been shown that \( L(N,p) \approx \frac{1}{K} N^{1/d} f(pKN) \), where \( d \) is the dimension of the lattice, \( f(x) = \text{const} \) if \( x \ll 1 \), and \( f(u) = \ln(u)/u \) if \( u \gg 1 \)

Clustering Coefficient & Degree Distribution for WS Model

- $C(p=0) = \frac{3(k - 2)}{4(k - 1)}$, $C(p=1) = k/N$
- $C(p) \approx C(p=0) \cdot (1-p)^3$, because we need to maintain three sides of the triangle.
- The degree $k_i$ of vertex $i$ can be written as $k_i = K/2 + c_1 + c_2$, where $K/2$ is the original edge of the unrewired end (since only single end of every edge is rewired), $c_1$ represents edges have been left in place (with prob=1-p) and $c_2$ represents the edges been rewired towards $i$, each with probability 1/N.
- The probability distribution of $c_1$ and $c_2$ are:
  - $P(c_1) = C_{c_1}^{K/2} (1-p)^{c_1} p^{K/2-c_1}$
  - $P(c_2) = C_{c_2}^{pNK/2} (1/N)^{c_2} (1-1/N) p^{pNK/2-c_2}$
  - $P(k) = \sum_{n=0}^{\min(K/2,K/2)} C_n^{K/2} (1-p)^n p^{K/2-n} \frac{(pK/2)^{(k-K/2-n)}}{(k-K/2-n)!} e^{-pK/2}$
Degree Distribution of Small-world Network

- Only one end of each edge is rewired
- Each node will have at least $K/2$ edges
- For $K>2$, no isolated nodes

$P(k)$

- Random graph (exact) with average degree

2009/9/29 SNA09, Modeling, Prof. Sd Lin
Giant Component Size of Power-law Graphs

\[ P(k) = Ck^{-\alpha} \Rightarrow \log(P(k)) = c - \alpha \log(k) \]

- The evolution of power-law graph using configuration model depends on \( \alpha \)
  - \( \alpha > 3.48 \): no giant component
  - \( \alpha < 3.48 \): there is almost surely a unique giant component
  - \( 2 < \alpha < 3.48 \): the second largest component is \( O(\log n) \), for any \( 2 \leq k < O(\log n) \), there is almost surely a component of size \( k \)
  - \( \alpha \sim 2 \): the second largest component is \( O(\log n/\log\log n) \)
  - \( 1 < \alpha < 2 \): the second largest component is \( O(1) \), and the graph is almost surely NOT connected
  - \( 0 < \alpha < 1 \): the graph is almost surely connected


The number of connected components for each possible component sizes for a call graph of a typical day

- The giant component contains nearly all of the nodes
- The maximum size of the next largest component is indeed exponentially smaller than the size of the giant component
- Interestingly, the distribution of the number of components of size smaller than the giant component is nearly log-log linear

The data was compiled by AT&T lab from raw phone call records, and the degree of the graph follows power law

The corresponding power-law distribution:
The number of connected components vs. component size for a collaboration network of power law

- A collaboration power-law graph with $\alpha=2.97$
- 337000 nodes (authors), 496000 edges (joint publication)
Outline

• **Random Graph**
  – Erdos-Renyi Model
  – Configuration Model

• **Scale-free Network**
  – Power-law distribution
  – Barabasi-Albert Model

• **Small-world Network**
  – Watts-Strogatz Model

• **Comparison of Network Models**
Component size distribution vs. Different degree distributions

• Let $\pi_s$ be the probability of a randomly chosen node belonging to a component of size $s$

• For the **Poisson** Random Graph

\[ p_k = e^{-c} \frac{c^k}{k!} \quad \pi_s = \frac{e^{-cs}(cs)^{s-1}}{s!} \]

• For the Graphs with **Exponential** distribution

\[ p_k = Ce^{-\lambda k} \]

\[ \pi_s \sim se^{-\mu s}, \text{ where } \mu = 2 \ln \left[ \frac{3}{2} (1 - e^{-\lambda}) \right] - \lambda \]

• For the **Power-law** Graphs

\[ p_k \propto k^{-\alpha} \]

\[ \pi_s = [1 - \ln 2]^{-1} \frac{(3s - 5)!}{(s - 1)! (2s - 2)!} s^{2 - 3s} \]

Component size distribution vs. Different degree distributions

- Poisson: $c=1.5$, Exponential: $\lambda=1$, Power-law: $\alpha=2.5$
- Each point is an average over 5000 networks of $10^6$ nodes
## Comparison of Network Models

<table>
<thead>
<tr>
<th></th>
<th>ER Model</th>
<th>Regular Lattice</th>
<th>WS Model</th>
<th>BA Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Path Length</strong></td>
<td>Short</td>
<td>Long</td>
<td>Short</td>
<td>Short</td>
</tr>
<tr>
<td></td>
<td>$l_{rand} \sim \log(N)/\log(z)$</td>
<td>$l_{lattice} \sim N/2K$</td>
<td>$l_{ws} \sim \log(N)$</td>
<td>$l_{sf} \sim \log(N)$</td>
</tr>
<tr>
<td><strong>Clustering Coefficient</strong></td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>$C_{rand} = z/n$</td>
<td>$C_{lattice} = 3/4$</td>
<td>$C_{ws}(p) = C(0)(1-p)^3$ independent of $N$</td>
<td>$C_{sf} = N^{-0.75}$</td>
</tr>
<tr>
<td><strong>Long Tail</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>$P(k) = z^k e^{-z}/k!$</td>
<td>$P(k) = k$</td>
<td>similar to random graph</td>
<td>$P(k) = Ck^{-\alpha}$</td>
</tr>
</tbody>
</table>

2009/9/29