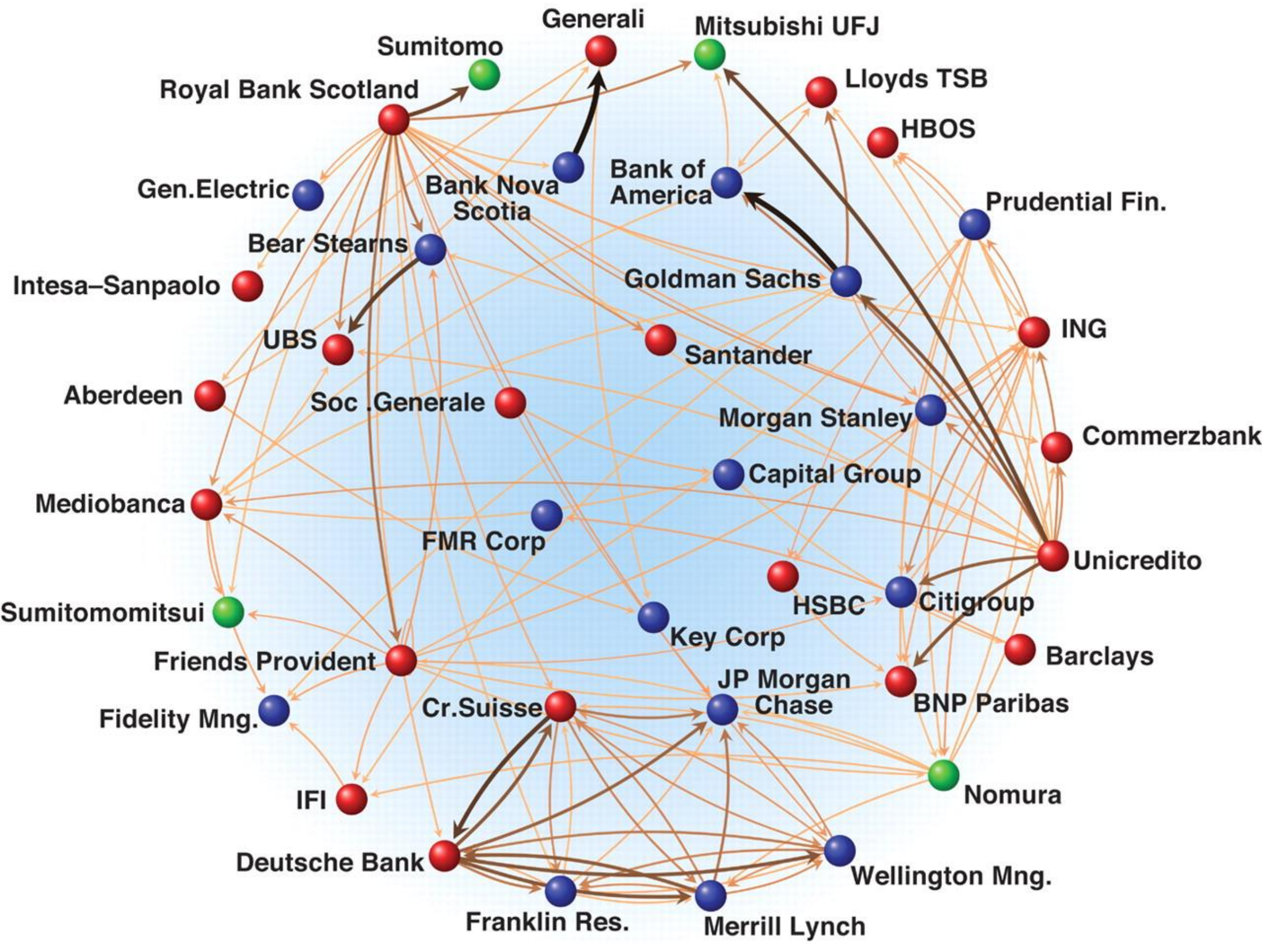
Business Network Analytics



F Schweitzer et al. Science 2009

Sep, 2017

Daning Hu
Department of Informatics
University of Zurich
Business Intelligence Research Group

Outline

- Network Topological Analysis

- Network Models
  - Random Networks
  - Small-World Networks
  - Scale-Free Networks

- Ref Book: Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)
  - http://www.amazon.com/Social-Network-Analysis-Applications-Structural/dp/0521387078

Network Topological Analysis

- Network topology is the arrangement of the various elements (links, nodes, etc). Essentially, it is the topological structure of a network.

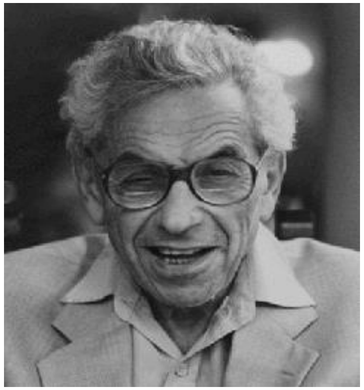
- How to model the **topology** of large-scale networks?

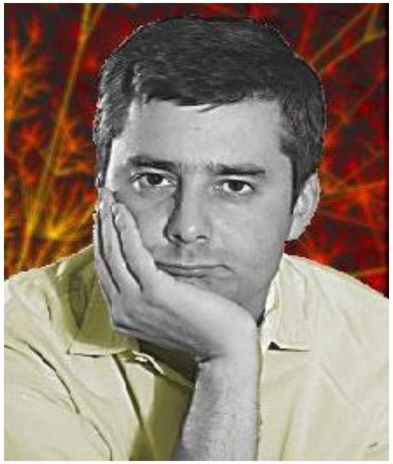
- What are the **organizing principles** underlying their topology?

- How does the topology of a network affect its **robustness** against errors and attacks?

Network Models

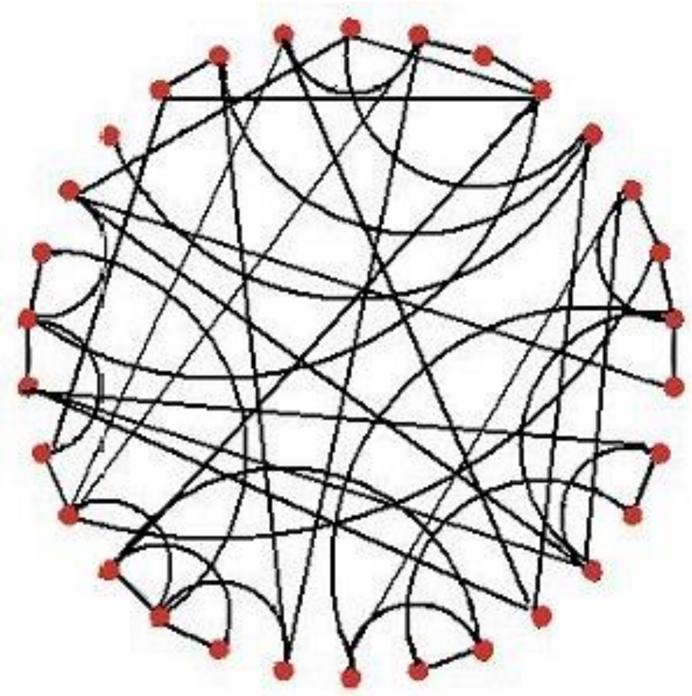
- Random graph model (Erdős & Rényi, 1959)
- Small-world model (Watts & Strogatz, 1998)
- Scale-free model (Barabasi & Alert, 1999)






Random Networks

- Erdős–Rényi Random Graph model is used for generating random networks in which links are set between nodes with equal probabilities
  - Starting with n isolated nodes and connecting each pair of nodes with probability p
  - As a result, all nodes have similar number of links (i.e., ***average degree, <k>***).



To Sum Up: Gnp

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

Path length: $O(\log n)$

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

Real (MSN) Networks vs. Gnp

- **Are real networks like random graphs?**
  - Giant connected component: ☺
  - Average path length: ☺
  - Clustering Coefficient: ☹
  - Degree Distribution: ☹
- **Problems with the random networks model:**
  - Degreed distribution differs from that of real networks
  - Giant component in most real network does NOT emerge through a phase transition
  - No local structure – clustering coefficient is too low
- **Most important: Are real networks random?**
  - The answer is simply: **NO!**

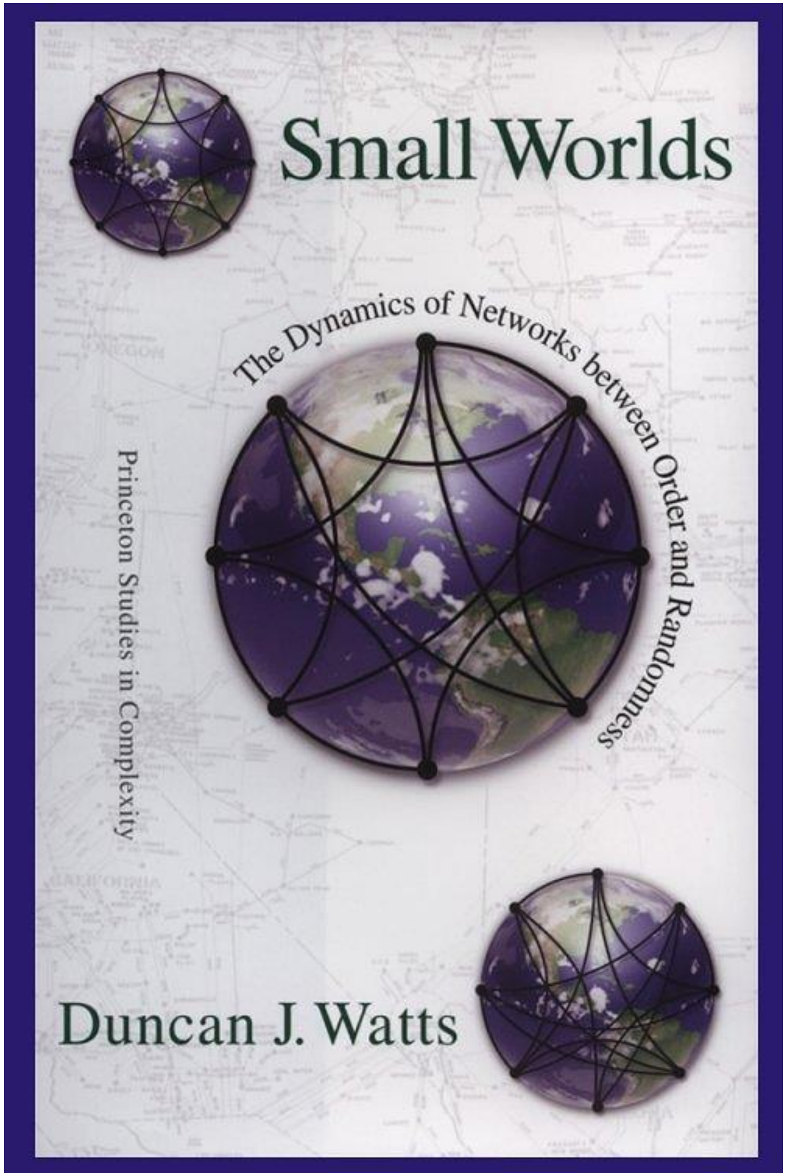
Small-World Network

Based on Milgram's (1967) famous work, the substantive point is that networks are structured such that even when most of our connections are local, any pair of people can be connected by a fairly small number of relational steps.

Works on 2 parameters:

1) The Clustering Coefficient (c) = average proportion of closed triangles

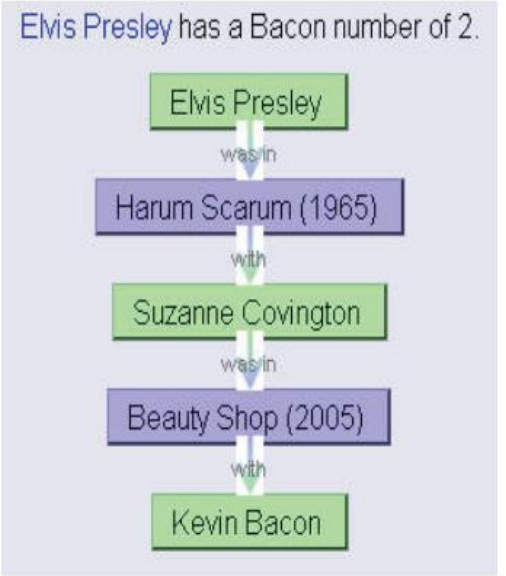
2) The average distance (L) separating nodes in the network



Six Degrees of Kevin Bacon

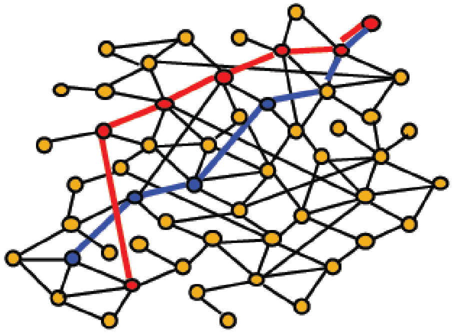
Origins of a small-world idea:

- **The Bacon number:**
  - Create a network of Hollywood actors
  - Connect two actors if they co-appeared in the movie
  - **Bacon number:** number of steps to Kevin Bacon
- As of Dec 2007, the highest (finite) Bacon number reported is 8
- Only approx. 12% of all actors cannot be linked to Bacon



The Small-World Experiment

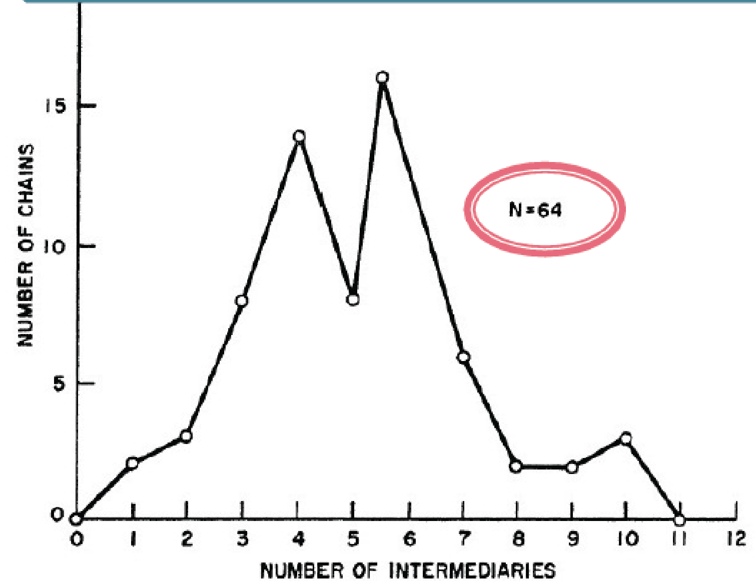
- What is the typical shortest path length between any two people?
  - Experiment on the global friendship network
    - Can’t measure, need to probe explicitly
- Small-world experiment [Milgram ’67]
  - Picked 300 people in Omaha, Nebraska and Wichita, Kansas
  - Ask them to get a letter to a stock-broker in Boston by passing it through friends
- How many steps did it take?



The Small-World Experiment

- **64 chains completed:**
  (i.e., 64 letters reached the target)
  - It took 6.2 steps on the average, thus **“6 degrees of separation”**
- **Further observations:**
  - People who owned stock had shorter paths to the stockbroker than random people: 5.4 vs. 6.7
  - People from the Boston area have even closer paths: 4.4

Milgram’s small world experiment

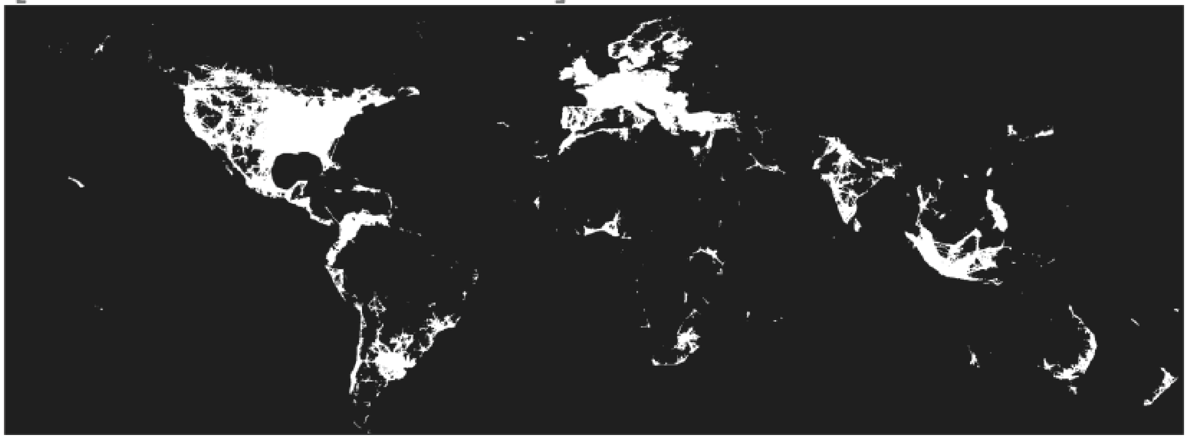


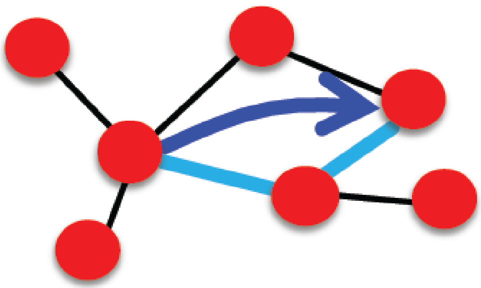
Columbia Small-World Experiment

- **In 2003 Dodds, Muhamad and Watts performed the experiment using e-mail:**
  - 18 targets of various backgrounds
  - 24,000 first steps (~1,500 per target)
  - 65% dropout per step
  - 384 chains completed (1.5%)

Six Degrees

- **Assume each human is connected to 100 other people**
  Then:
  - Step 1: reach 100 people
  - Step 2: reach 100*100 = 10,000 people
  - Step 3: reach 100*100*100 = 1,000,000 people
  - Step 4: reach 100*100*100*100 = 100M people
  - **In 5 steps we can reach 10 billion people**
- **What's wrong here?**
  - **92% of new FB friendships are to a friend-of-a-friend** [Backstom-Leskovec '11]





Clustering Implies Link Locality

- **MSN network has 7 orders of magnitude larger clustering than the corresponding G_{np}!**
- **Other examples:**

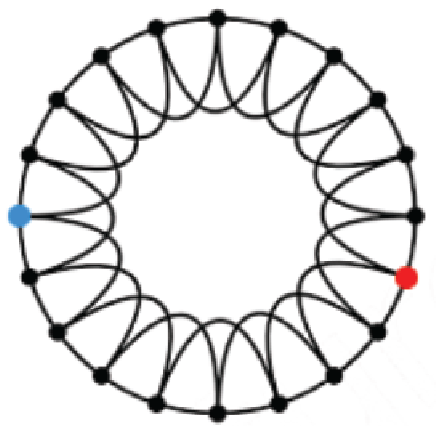
Actor Collaborations (IMDB): N = 225,226 nodes, avg. degree $\overline{k}$ = 61
Electrical power grid: N = 4,941 nodes, $\overline{k}$ = 2.67
Network of neurons: N = 282 nodes, $\overline{k}$ = 14

Network	h_{actual}	h_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power Grid	18.70	12.40	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

h ... Average shortest path length
C ... Average clustering coefficient
"actual" … real network
"random" … random graph with same avg. degree

Small-World: How?

- **Could a network with high clustering be at the same time a small world?**
  - How can we at the same time have **high clustering** and **small diameter?**



High clustering
High diameter

Low clustering
Low diameter

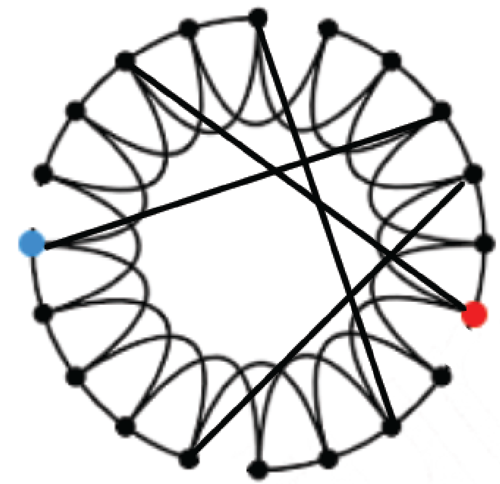
  - Clustering implies edge "locality"
  - Randomness enables "shortcuts"

Solution: The Small-World Model

Small-world Model [Watts-Strogatz ‘98]

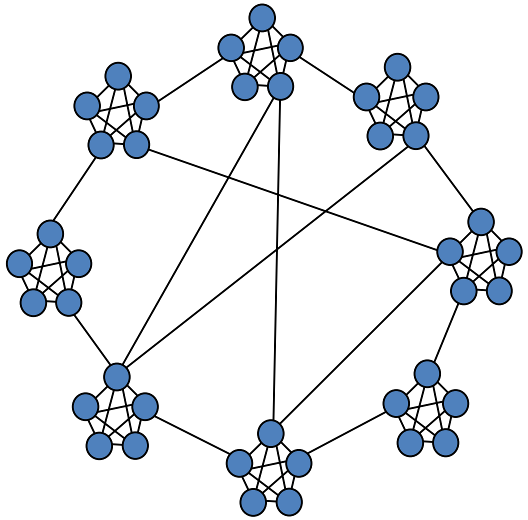
Two components to the model:

- **(1) Start with a low-dimensional regular lattice**
  - (In our case we using a ring as a lattice)
  - Has high clustering coefficient
- **Now introduce randomness (“shortcuts”)**
- **(2) Rewire:**
  - Add/remove edges to create shortcuts to join remote parts of the lattice
  - For each edge with prob. p move the other end to a random node

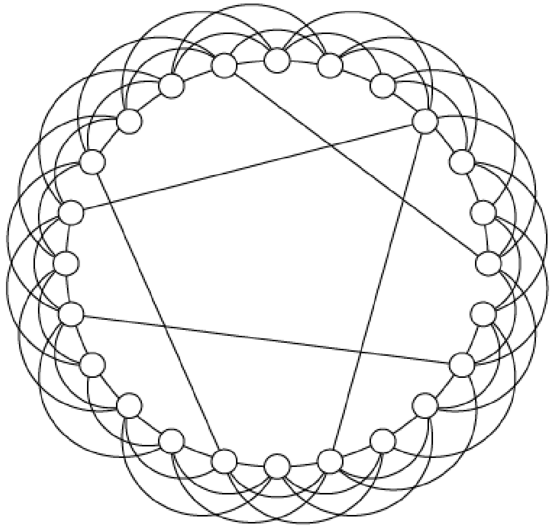


Small-World Network

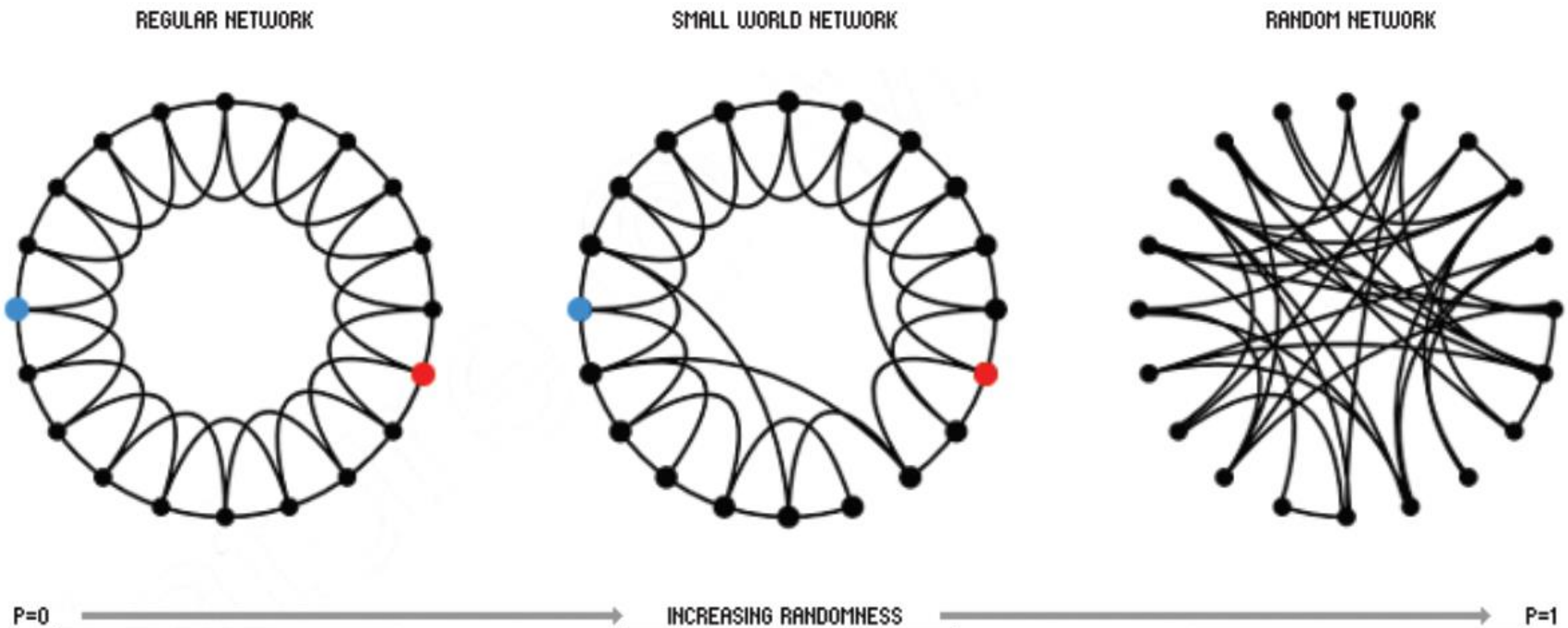
In a highly clustered, ordered network, a single random connection will create a shortcut that lowers *diameter* dramatically



Watts demonstrates that small world properties can occur in graphs with a surprisingly small number of shortcuts



The Small-World Model



High clustering High diameter	High clustering Low diameter	Low clustering Low diameter
$h = \frac{N}{2\bar{k}}$ $\quad C = \frac{3}{4}$		$h = \frac{\log N}{\log \alpha}$ $\quad C = \frac{\bar{k}}{N}$

Rewiring allows us to "interpolate" between
a regular lattice and a random graph

Small-World: Summary

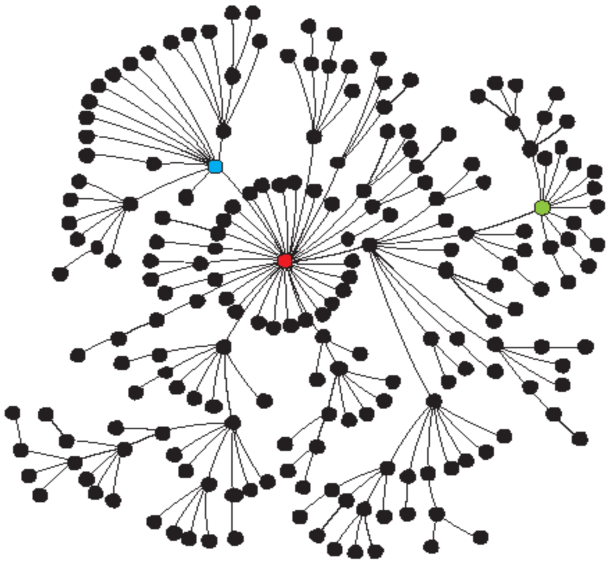
- **Could a network with high clustering be at the same time a small world?**
  - Yes! You don't need more than a few random links
- **The Watts Strogatz Model:**
  - Provides insight on the interplay between clustering and the small-world
  - Captures the structure of many realistic networks
  - Accounts for the high clustering of real networks

Scale-Free (SF) Networks: Barabási–Albert (BA) Model

- “Scale free” means there is no single characterizing degree in the network

- **Growth**:
  - starting with a small number (n_0) of nodes, at every time step, we add a new node with $m(<=n_0)$ links that connect the new node to m different nodes already present in the system

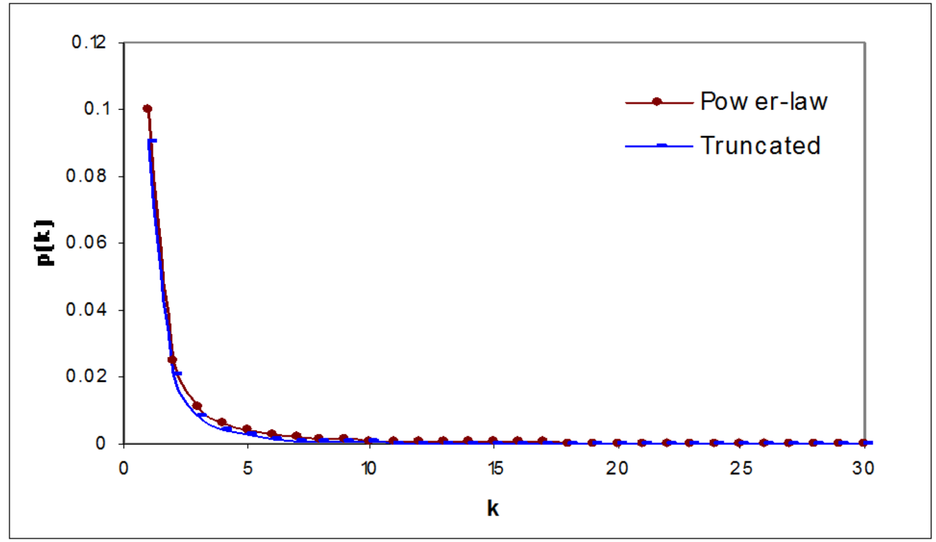
- **Preferential attachment**:
  - When choosing the nodes to which the new node will be connected to node i depends on its degree k_i



Scale-Free Networks (Cont’d)

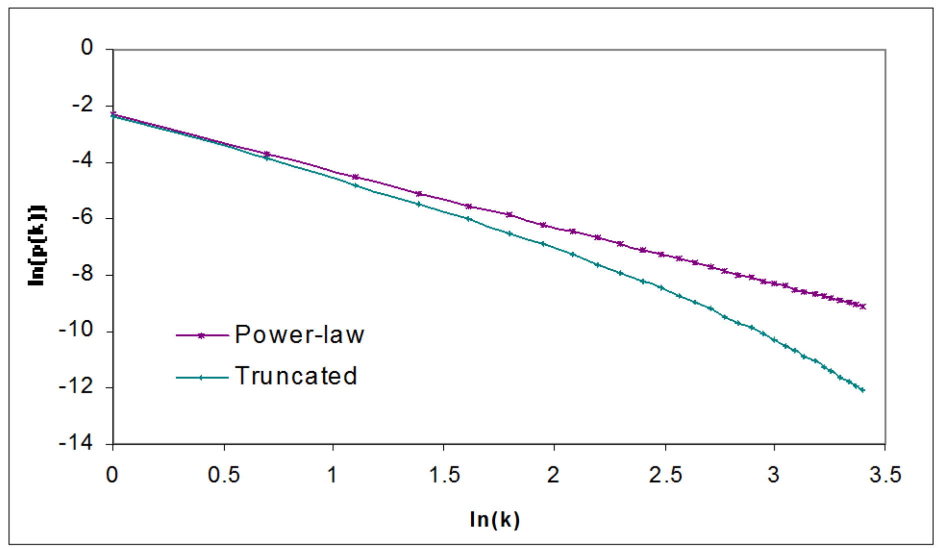
- The degree of scale-free networks follows **power-law distribution** with a flat tail for large *k*

$$p(k) \sim k^{-\gamma}$$



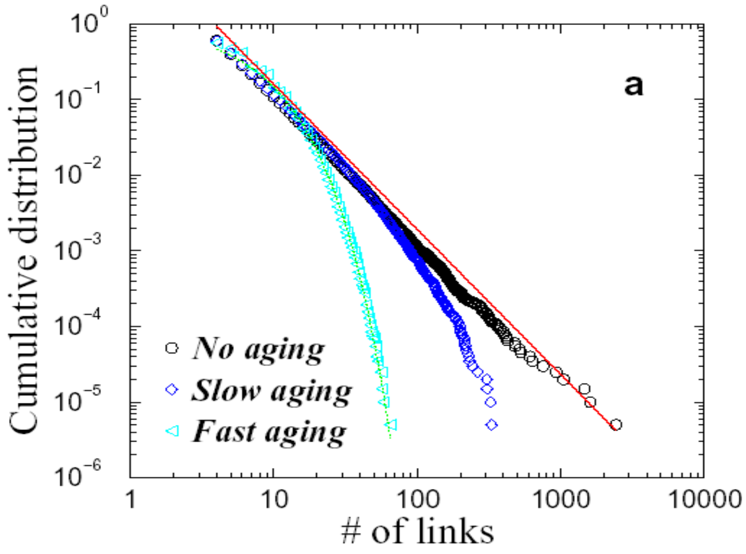
- Truncated power-law distribution deviates at the tail

$$p(k) \sim k^{-\gamma} e^{-\frac{k}{\kappa}}$$

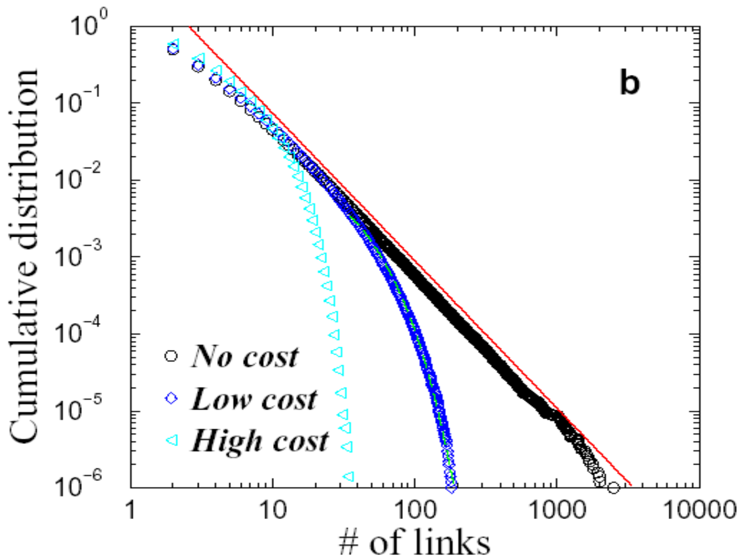


Evolution of SF Networks

- The emergence of scale-free network is due to
  - Growth effect: new nodes are added to the network
  - Preferential attachment effect (Rich-get-richer effect): new nodes prefer to attach to "popular" nodes



- The emergence of truncated SF network is caused by some constraints on the maximum number of links a node can have such as (Amaral, Scala et al. 2000)
  - Aging effect: some old nodes may stop receiving links over time
  - Cost effect: as maintaining links induces costs, nodes cannot receive an unlimited number of links



Network Analysis: Topology Analysis

Topology	Average Path Length (L)	Clustering Coefficient (CC)	Degree Distribution ($P(k)$)
Random Graph	$L_{rand} \sim \frac{\ln N}{\ln\langle k\rangle}$	$CC_{rand} = \frac{\langle k\rangle}{N}$	Poisson Dist.: $P(k) \approx e^{-\langle k\rangle} \frac{\langle k\rangle^k}{k!}$
Small World (Watts & Strogatz, 1998)	$\boldsymbol{L_{sw} \leq L_{rand}}$	$\boldsymbol{CC_{sw} >> CC_{rand}}$	Similar to random graph
Scale-Free network	$L_{SF} \leq L_{rand}$		**Power-law Distribution:** $\boldsymbol{P(k) \sim k^{-\gamma}}$

$\langle k\rangle$: Average degree

Empirical Results from Real-World Networks

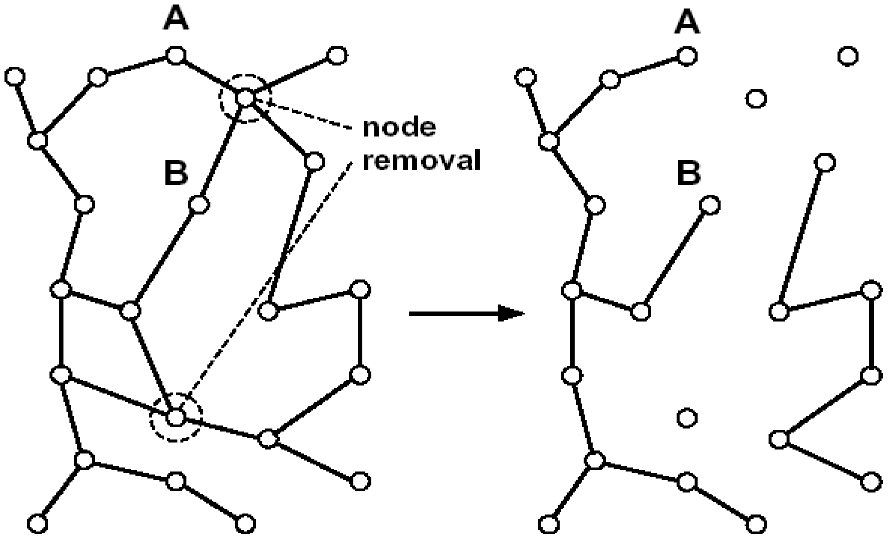
Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}	ℓ_{real}	ℓ_{rand}	ℓ_{pow}	Reference
WWW	$325,729$	4.51	900	2.45	2.1	11.2	8.32	4.77	Albert, Jeong, Barabási 1999
WWW	4×10^7	7		2.38	2.1				Kumar *et al.* 1999
WWW	2×10^8	7.5	$4,000$	2.72	2.1	16	8.85	7.61	Broder *et al.* 2000
WWW, site	$260,000$				1.94				Huberman, Adamic 2000
Internet, domain*	$3,015$ - $4,389$	3.42 - 3.76	$30 - 40$	2.1 - 2.2	2.1 - 2.2	4	6.3	5.2	Faloutsos 1999
Internet, router*	$3,888$	2.57	30	2.48	2.48	12.15	8.75	7.67	Faloutsos 1999
Internet, router*	$150,000$	2.66	60	2.4	2.4	11	12.8	7.47	Govindan 2000
Movie actors*	$212,250$	28.78	900	2.3	2.3	4.54	3.65	4.01	Barabási, Albert 1999
Coauthors, SPIRES*	$56,627$	173	$1,100$	1.2	1.2	4	2.12	1.95	Newman 2001b,c
Coauthors, neuro.*	$209,293$	11.54	400	2.1	2.1	6	5.01	3.86	Barabási *et al.* 2001
Coauthors, math*	$70,975$	3.9	120	2.5	2.5	9.5	8.2	6.53	Barabási *et al.* 2001
Sexual contacts*	2810			3.4	3.4				Liljeros *et al.* 2001
Metabolic, E. coli	778	7.4	110	2.2	2.2	3.2	3.32	2.89	Jeong *et al.* 2000
Protein, S. cerev.*	1870	2.39		2.4	2.4				Mason *et al.* 2000
Ythan estuary*	134	8.7	35	1.05	1.05	2.43	2.26	1.71	Montoya, Solé 2000
Silwood park*	154	4.75	27	1.13	1.13	3.4	3.23	2	Montoya, Solé 2000
Citation	$783,339$	8.57			3				Redner 1998
Phone-call	53×10^6	3.16		2.1	2.1				Aiello *et al.* 2000
Words, cooccurence*	$460,902$	70.13		2.7	2.7				Cancho, Solé 2001
Words, synonyms*	$22,311$	13.48		2.8	2.8				Yook *et al.* 2001

Implications of Network Modeling

- The two new models of networks have important implications to many applications, e.g.,
  - The 19 degrees of separation on the WWW implies that on average, a user can navigate from an arbitrary web page to another randomly selected page within 19 clicks, even though the WWW consists of millions of pages. Even if the web increase by 10 times in the next few years, the average path length increases only marginally from 19 to 21! (Albert, Jeong, & Barabási, 1999)
  - The small-world properties of metabolic networks in cell implies that cell functions are modulized and localized
- The ubiquity of SF networks lead to a conjecture that complex systems are governed by the same self-organizing principle(s).

Robustness Testing

- How will the topology of a network be affected if some nodes are removed from the network?
- How will random node removal (**failure**) and targeted node removal (**attack** targeting hubs) affect
  - ***S***: the fraction of nodes in the largest component
  - ***L***: the average path length of the largest component



Robustness Testing (Cont.)

- SF networks are **more robust against failures** than random networks due to its skewed degree distribution
- SF networks are **more vulnerable to attacks** than random networks, again, due to its skewed degree distribution
- The power-law degree distribution becomes the Achilles' Heel of SF networks