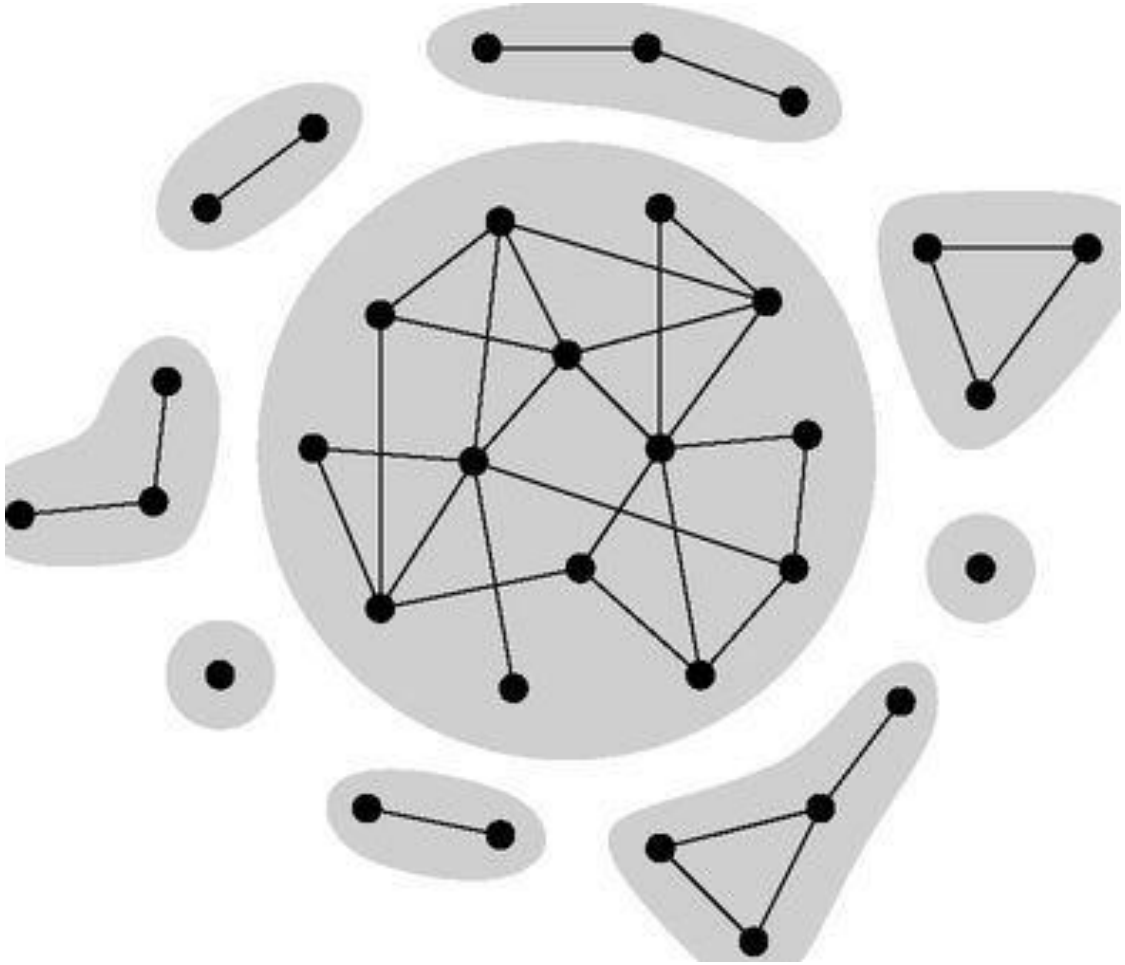
Generic Features of Real-World Networks

The More Things Change, the More They Stay the Same

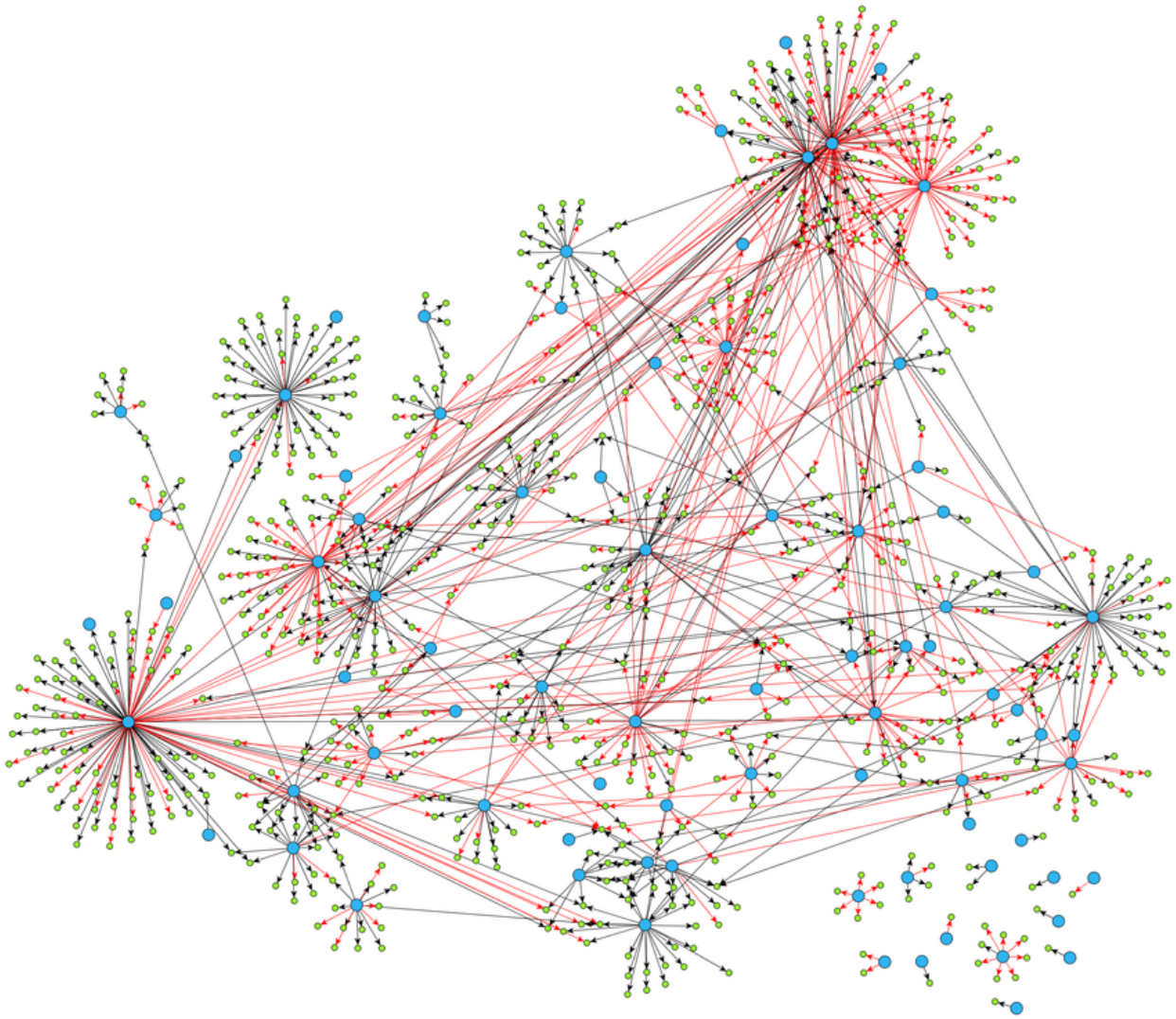
Although the nodes and edges of real networks vary greatly from one to the next, some features show up over and over again. These generic features are the topic of this chapter.

Giant Components



Gene Regulatory Network

Most nodes are part of a single, large component. This is called a giant component.

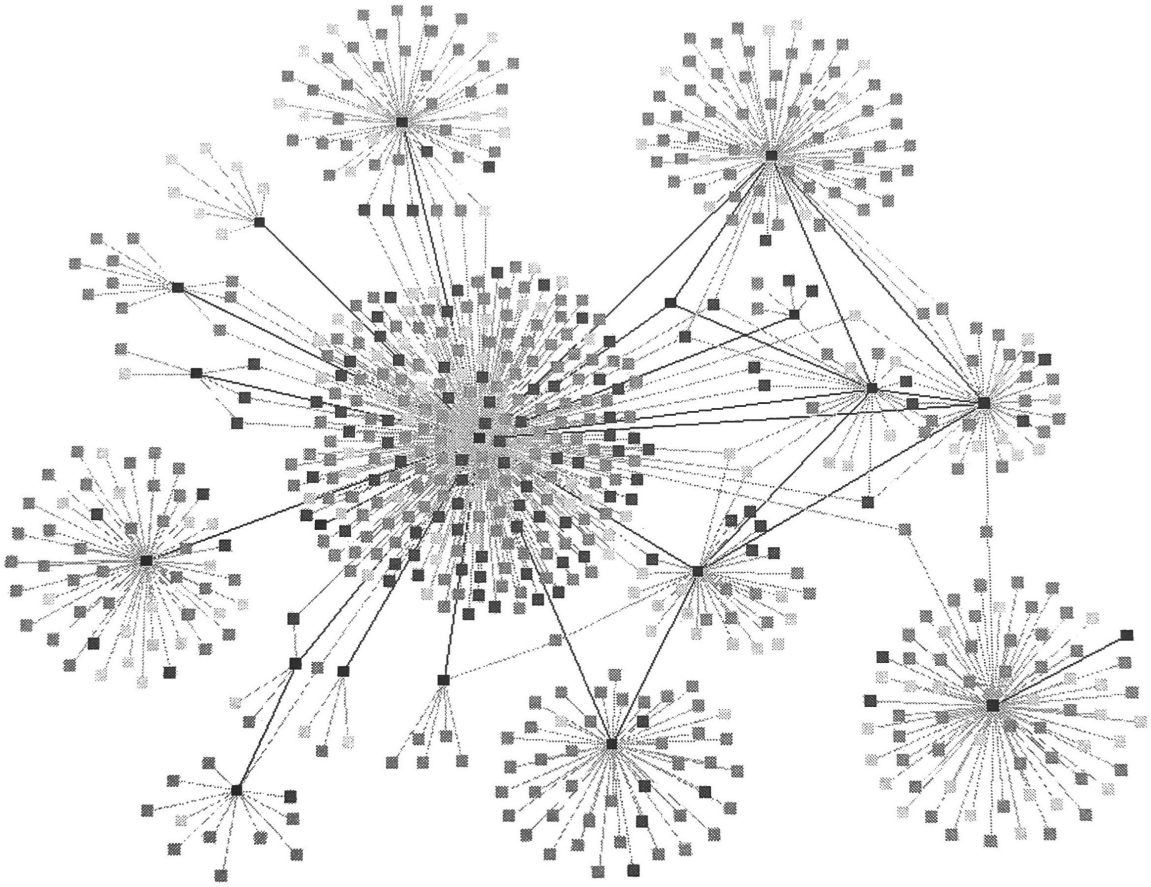


Washington DC Subway Map

Each subway station can be reached by any other. The giant component is the entire network.

Spread of Tuburculosis

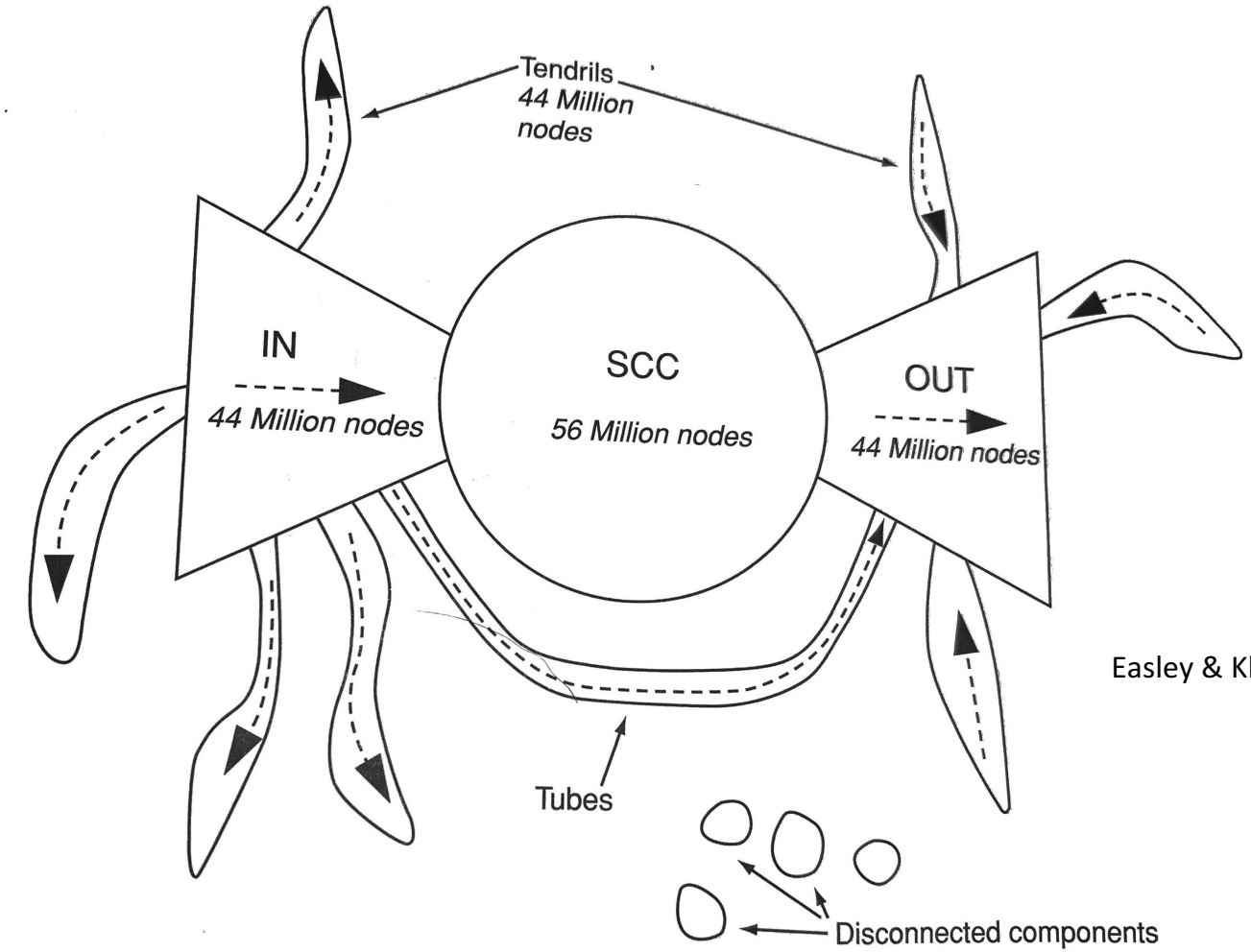
Nodes are people, an edge means that one person infected another.
Again, the giant component is the entire network.



What does this imply about the spread of the disease in this community?

Bowtie Structure of the WWW

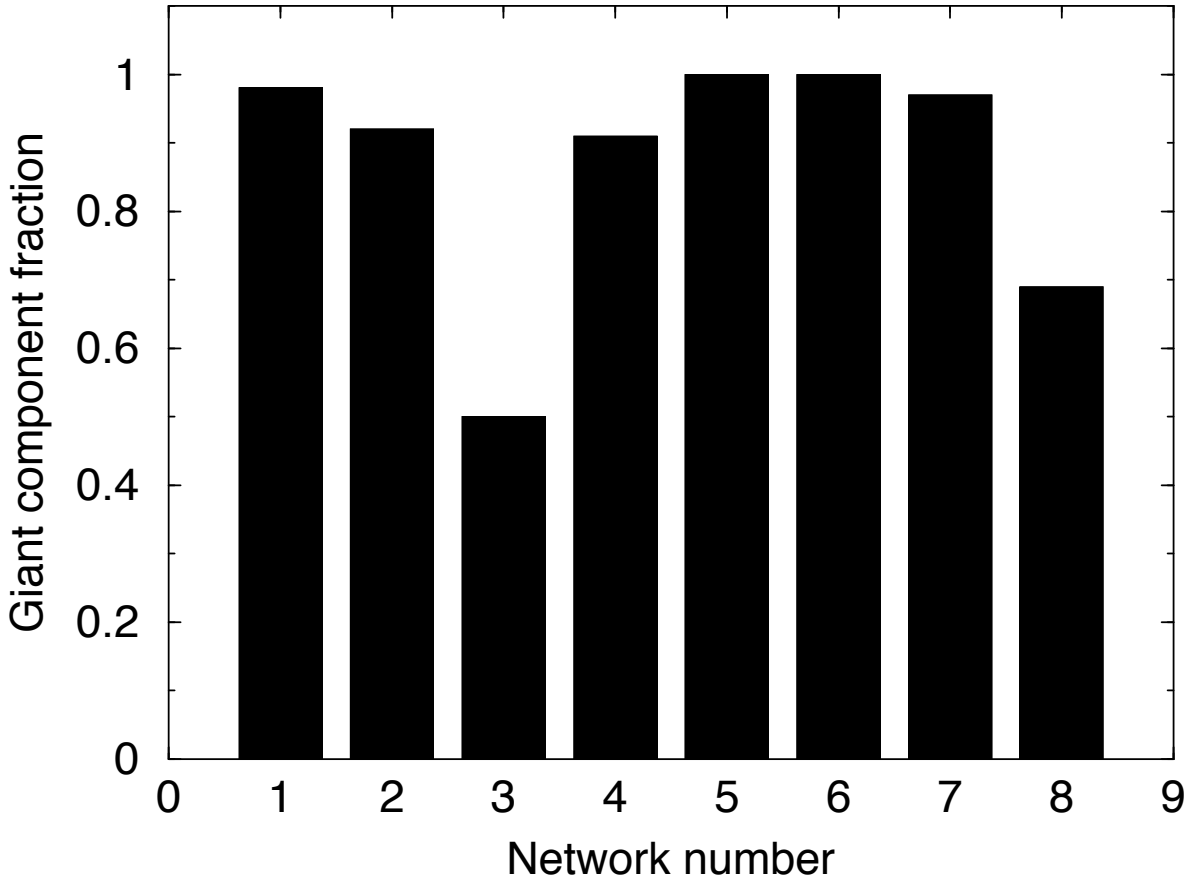
Data from 1999, but structure still accurate today.



Easley & Kleinberg, 2010

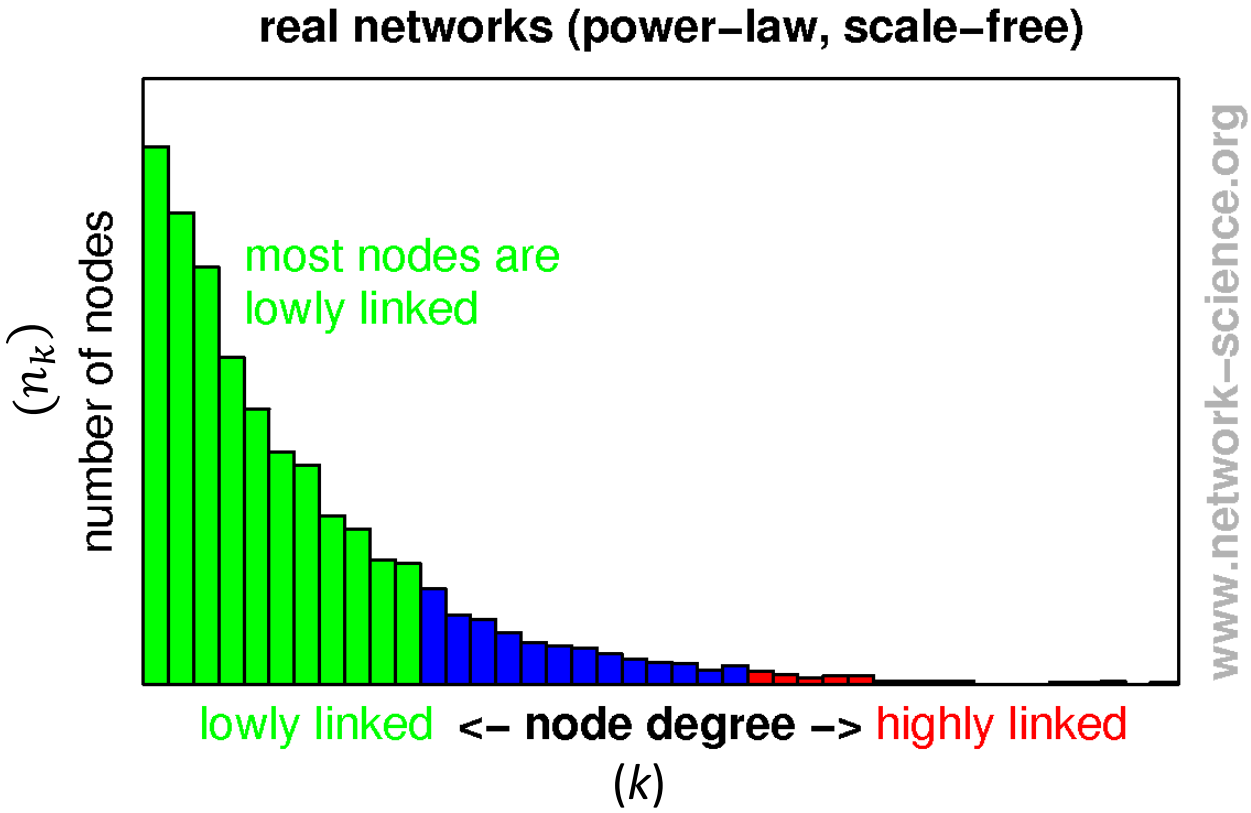
SCC=Strong Connected Component
IN=Web pages linking to the SCC, but not linked from the SCC
OUT=Web pages linked from the SCC, but not linking to the SCC

Most Real-Life Networks Have a Single Giant Component



1. Film actors
2. Math coauthorship
3. Student dating
4. WWW (weakly connected)
5. Internet
6. Power grid
7. Metabolic network
8. Protein interactions

In Real-World Networks, Most Nodes Have Low Degree

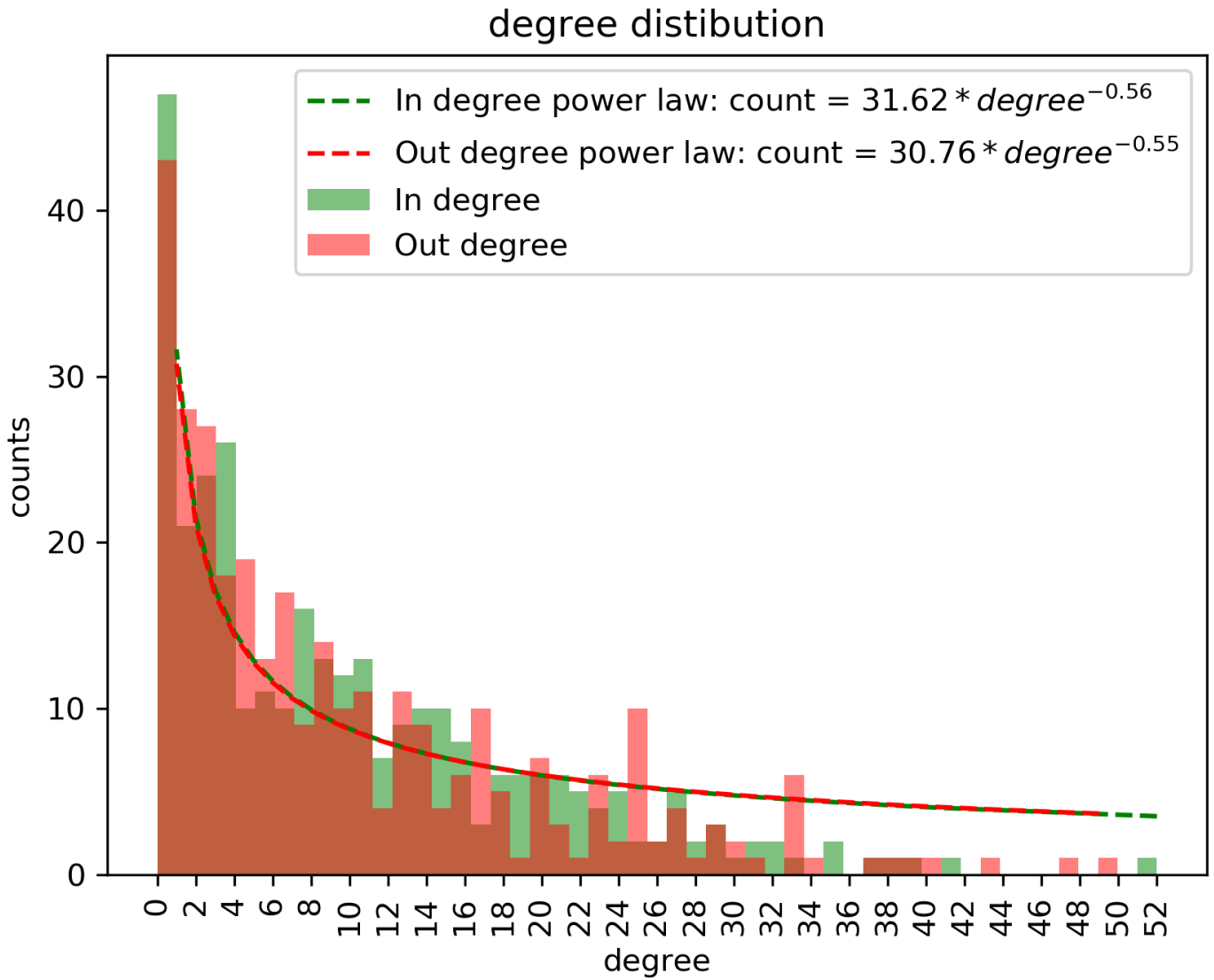


In real networks, most nodes have a low degree, but a few hubs have a high degree. The histogram of the degree distribution follows a power function:

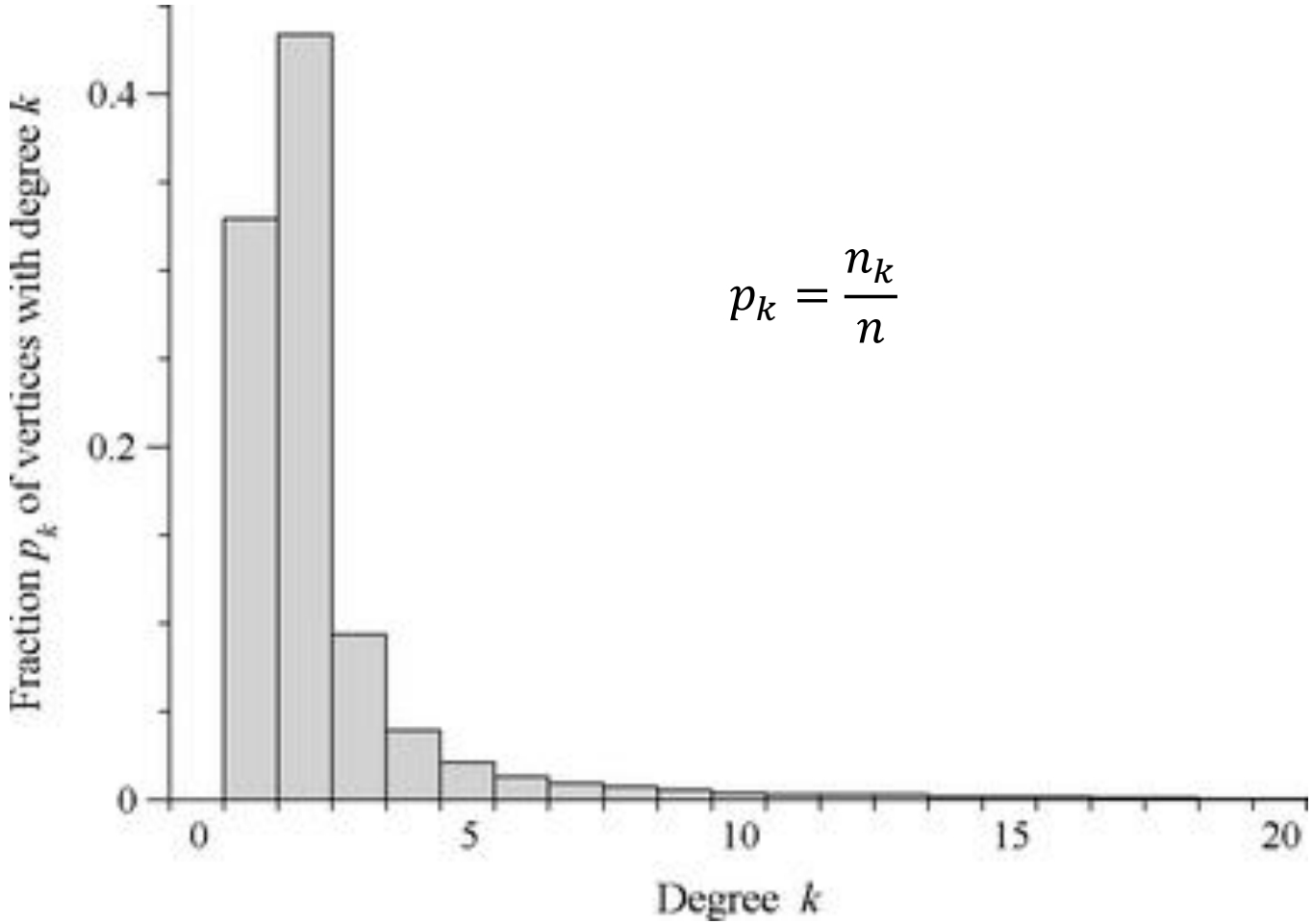
$n_k = Ck^{-\alpha}$, where k is degree and $\alpha > 0$. Typically, $2 < \alpha < 3$ for a real network.

Power Law Degree Distribution

Gene transcription network



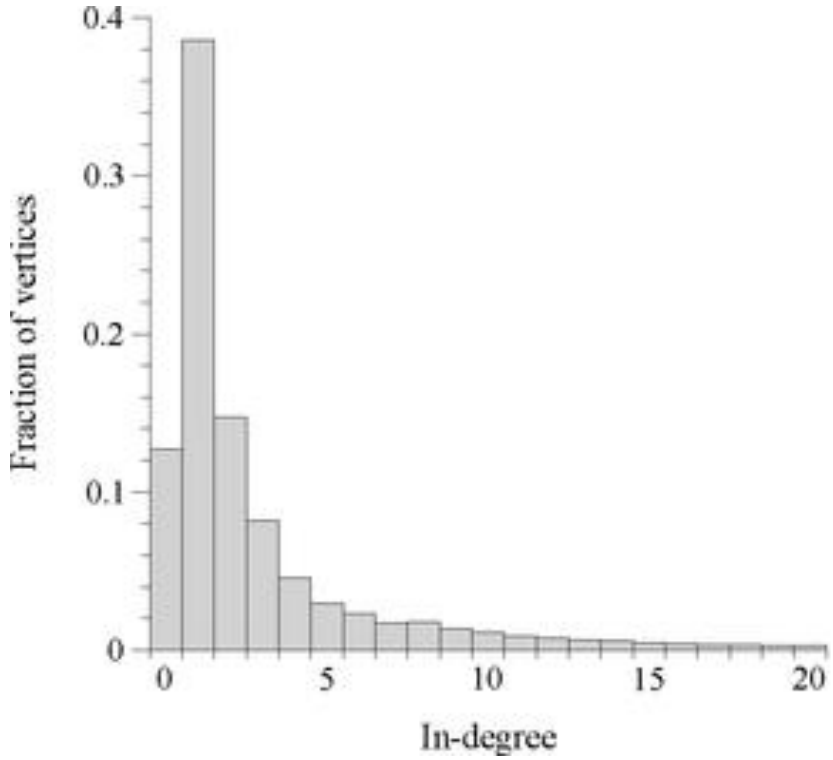
Degree Distribution of the Internet

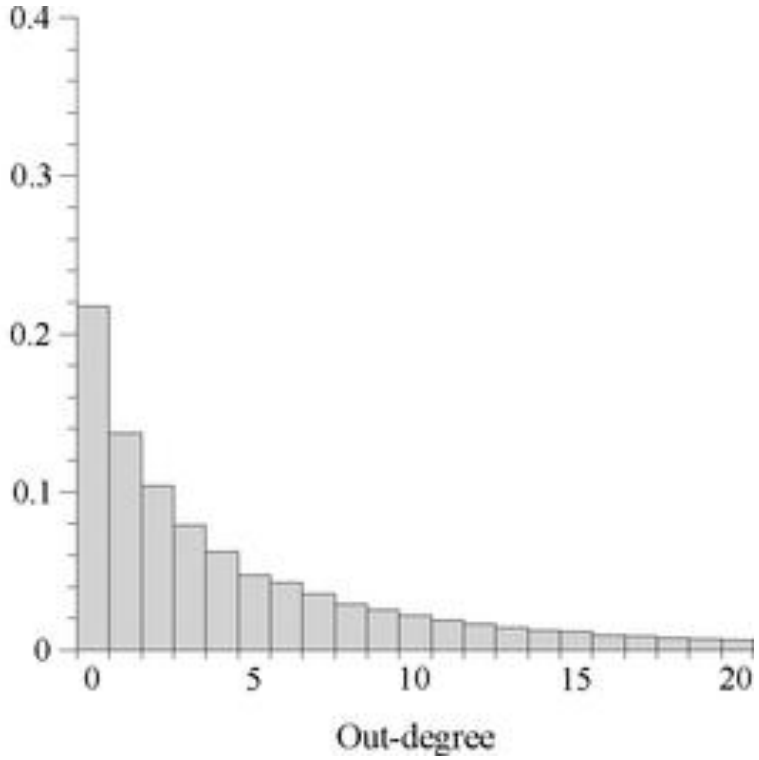


$$p_k = \frac{n_k}{n}$$

Note that for nodes of very low degree the power law does not hold.
This is common in real-life networks.

In-Degree and Out-Degree Distributions of the WWW





A Power Law Distribution Looks Linear in log-log Plots

Why linear? Linearity in a log-log plot means that $\ln p_k = -\alpha \ln k + C$, with $\alpha > 0$ and $C > 0$.

Exponentiating both sides, $p_k = e^{(-\alpha \ln k + C)}$, which simplifies to $p_k = Ck^{-\alpha}$.

Networks that have a power-law distribution are called scale-free networks. To see why, scale *k* by any number γ. Then $p_{\gamma k} = C(\gamma k)^{-\alpha} = C\gamma^{-\alpha}k^{-\alpha}$, or

$$p_{\gamma k} = \gamma^{-\alpha} p_k$$

That is, power law functions with the same exponent α are all just scaled versions of one another.

In contrast, an exponential function is not scale-free, since if $p_k = Ce^{-k}$, then

$$p_{\gamma k} = Ce^{-\gamma k} \neq e^{-\gamma} p_k$$

Scale-Free Networks are Heterogeneous

The power law degree distribution is characterized by many low-degree nodes and a smaller number of hubs. This heterogeneity can be quantified with a heterogeneity parameter κ.

If k_i is the degree of node i, then the average square degree of the network is $\langle k^2 \rangle = \frac{k_1^2 + k_2^2 + \cdots + k_N^2}{N} = \frac{\sum_{i=1}^{N} k_i^2}{N}$.

The heterogeneity parameter κ is defined as:

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$$

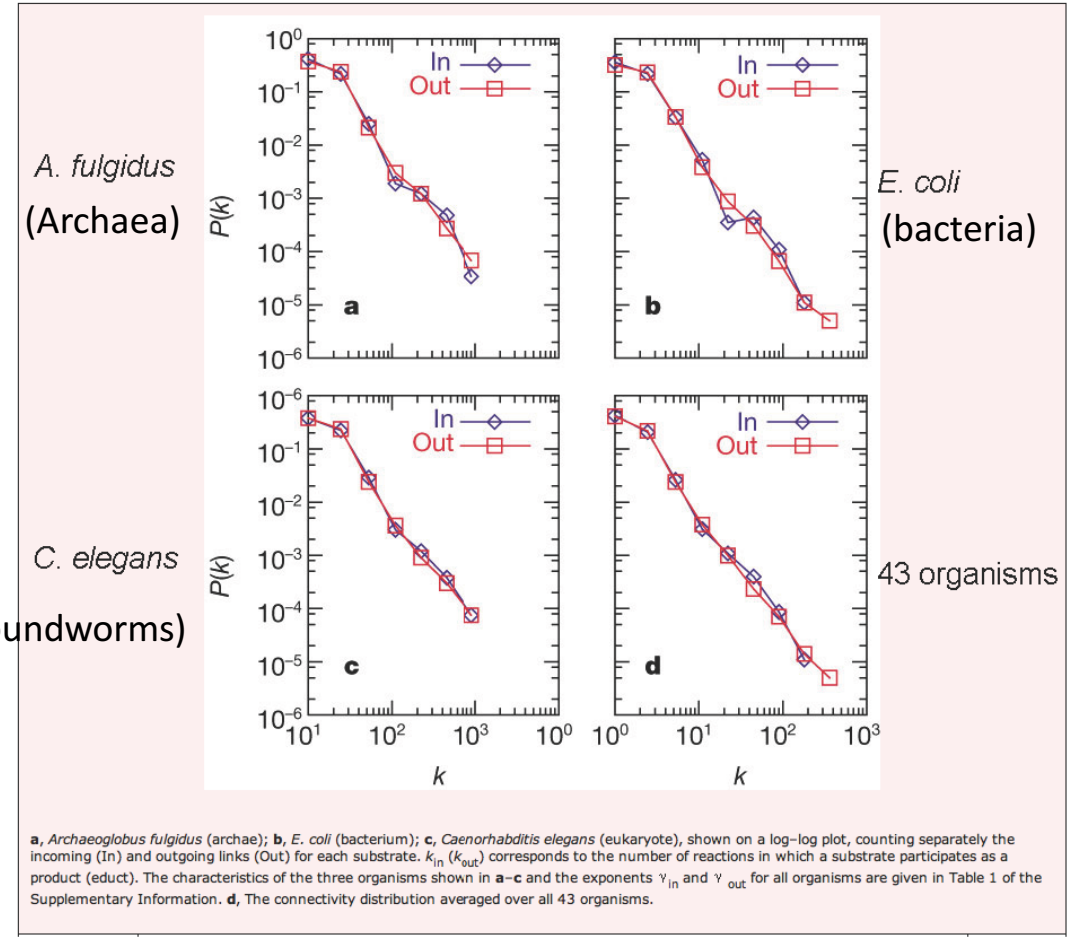
If all nodes have the same degree, then $\kappa = 1$. With degree heterogeneity $\kappa > 1$, and it gets larger for greater heterogeneity.

Real-Life Examples of Heterogeneous Networks

The power law degree distribution is characterized by many low-degree nodes and a smaller number of hubs. This heterogeneity can be quantified with a heterogeneity parameter κ.

Network	Heterogeneity parameter
Facebook Northwestern Univ.	1.8
C. elegans brain	2.7
US air transportation	5.3
Wikipedia math	38.2

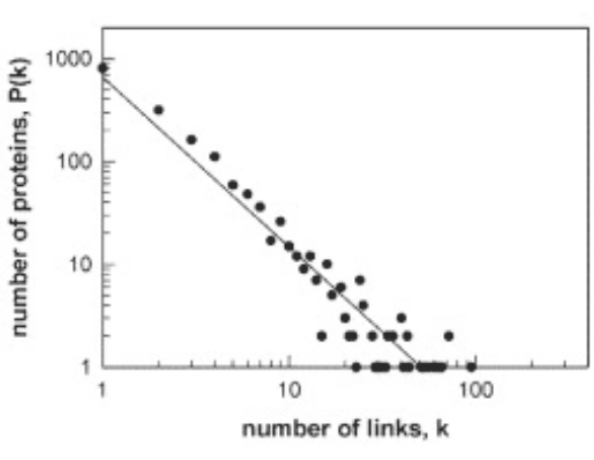
Scale-Free Metabolic Networks

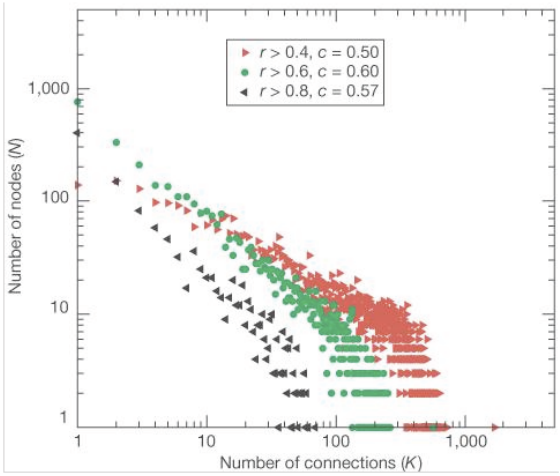


a, *Archaeoglobus fulgidus* (archae); **b**, *E. coli* (bacterium); **c**, *Caenorhabditis elegans* (eukaryote), shown on a log–log plot, counting separately the incoming (In) and outgoing links (Out) for each substrate. k_{in} (k_{out}) corresponds to the number of reactions in which a substrate participates as a product (educt). The characteristics of the three organisms shown in **a**–**c** and the exponents γ_{in} and γ_{out} for all organisms are given in Table 1 of the Supplementary Information. **d**, The connectivity distribution averaged over all 43 organisms.

Jeong et al, Nature, 407:651, 2000

Scale-Free Signaling and Transcription Networks





Protein interaction network
H. sapiens

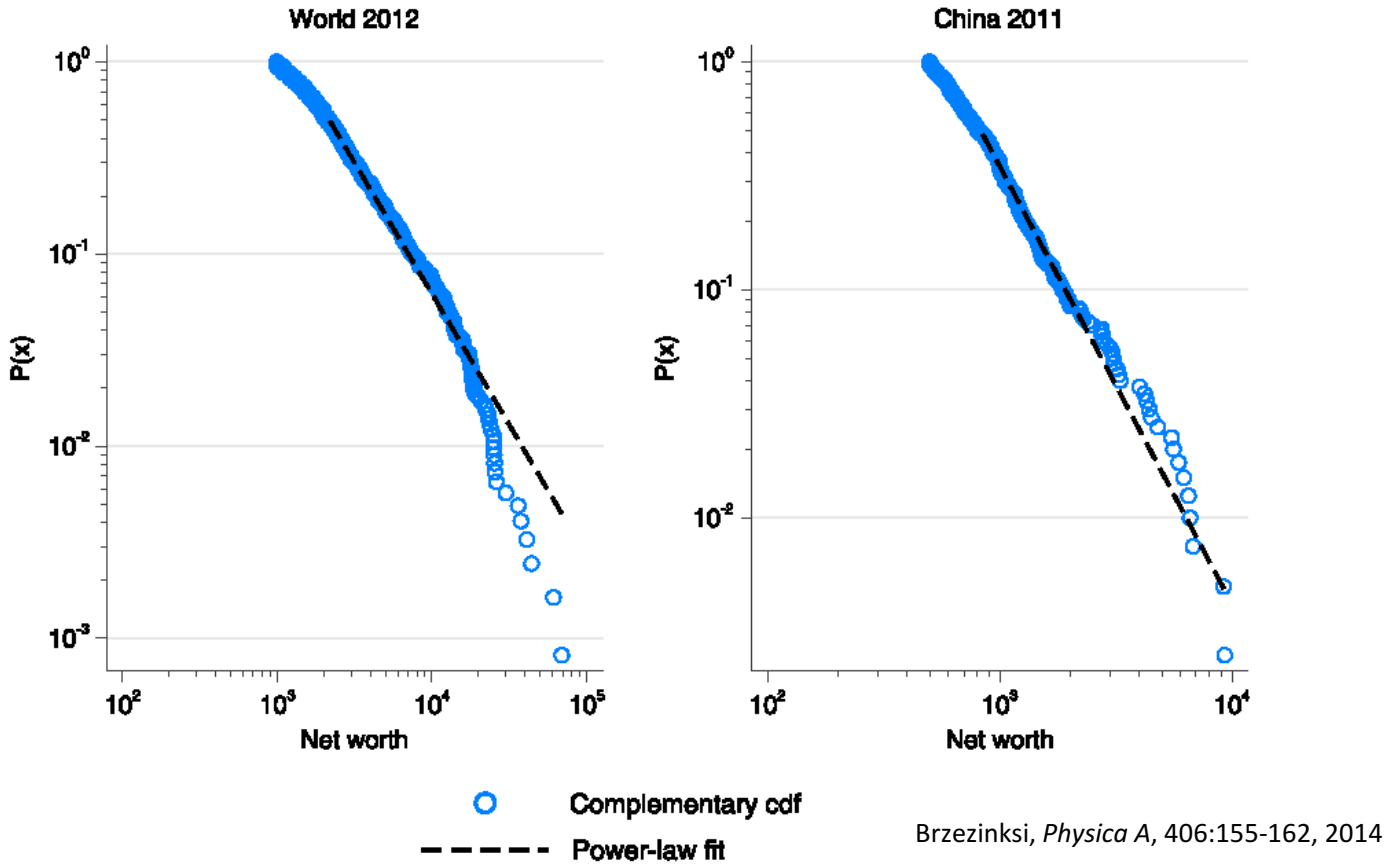
Stelzl et al. Cell, 122:957, 2005

Gene co-expression network
S. cerevisiae
(Brewers yeast)

Noort et al, EMBO Reports,5:280, 2004

Power Law Distributions Come Up Often in Applications

Histograms of wealth distribution.



Brzezinksi, *Physica A*, 406:155-162, 2014

Power-law exponents over different years range from 1.9 to 2.7

Some Other Examples

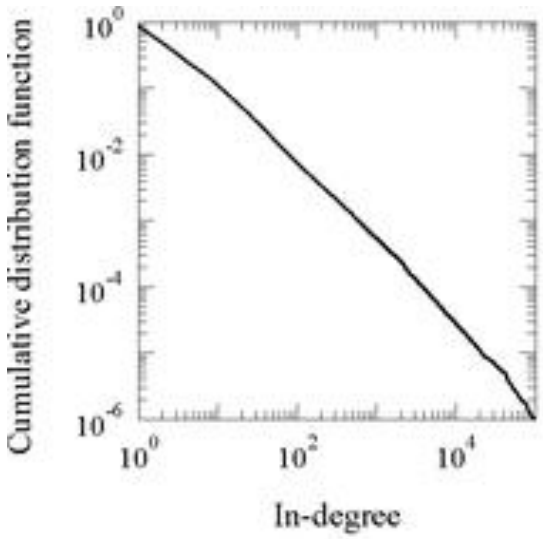
The fraction of telephone numbers that receive *k* calls per day is roughly proportional to k^{-2}.

The fraction of books that are bought by *k* people is roughly proportional to k^{-3}.

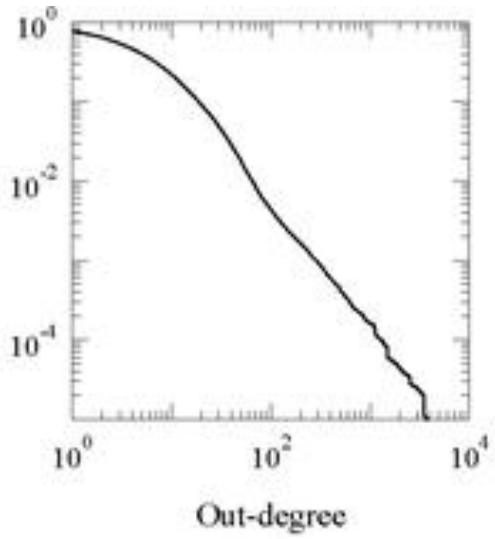
The fraction of scientific papers that receive *k* citations in total is roughly proportional to k^{-3}.

In general, power laws dominate in cases where the quantity being measured can be viewed as a type of popularity.

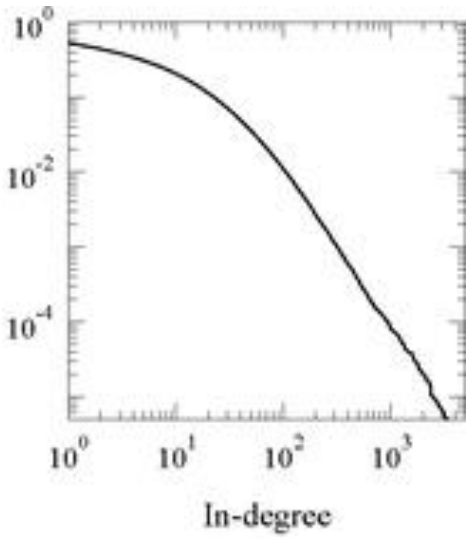
Power-Law Cumulative Distribution Function for In-Degrees and Out-Degrees



(a) World Wide Web

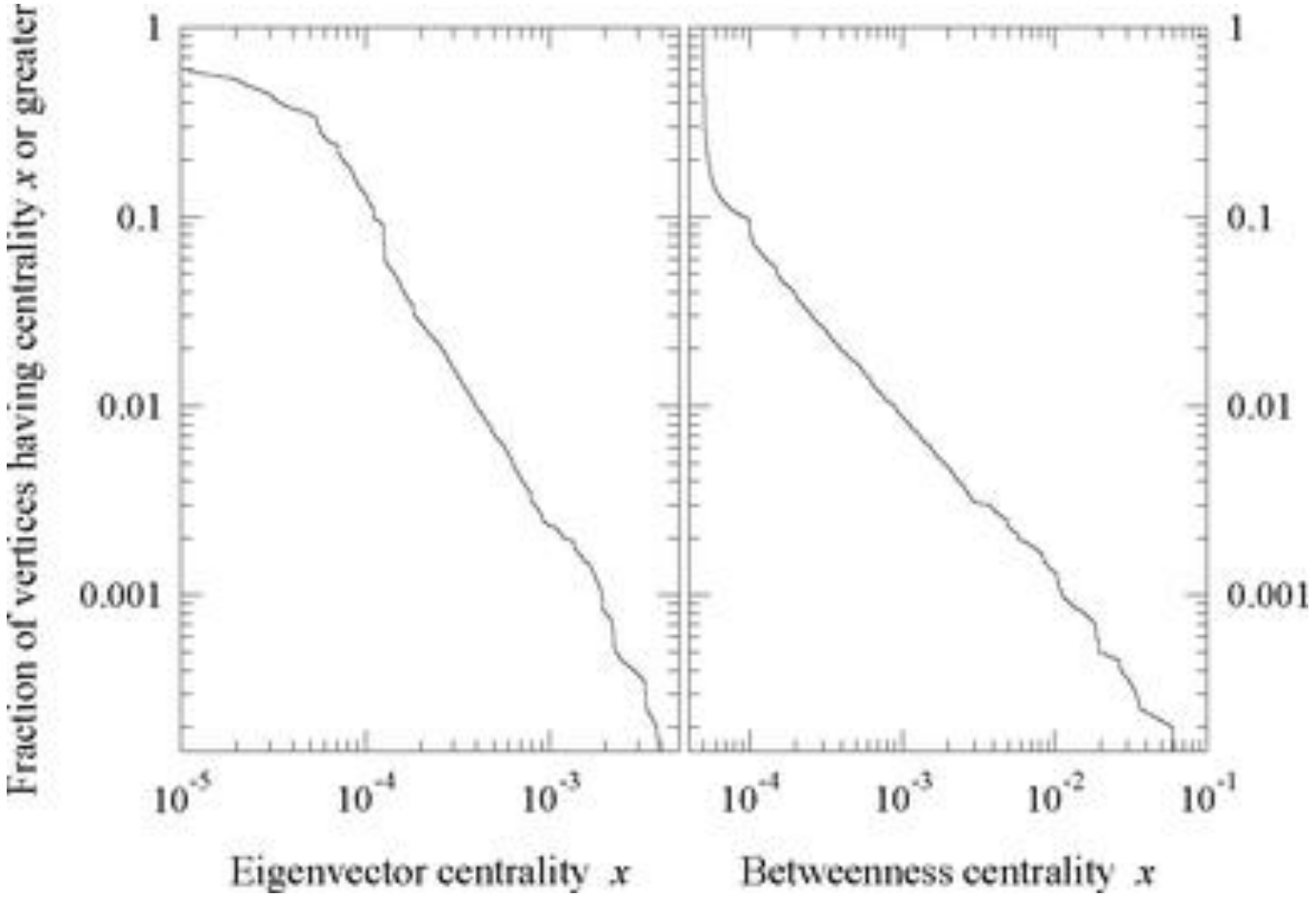


(b) World Wide Web



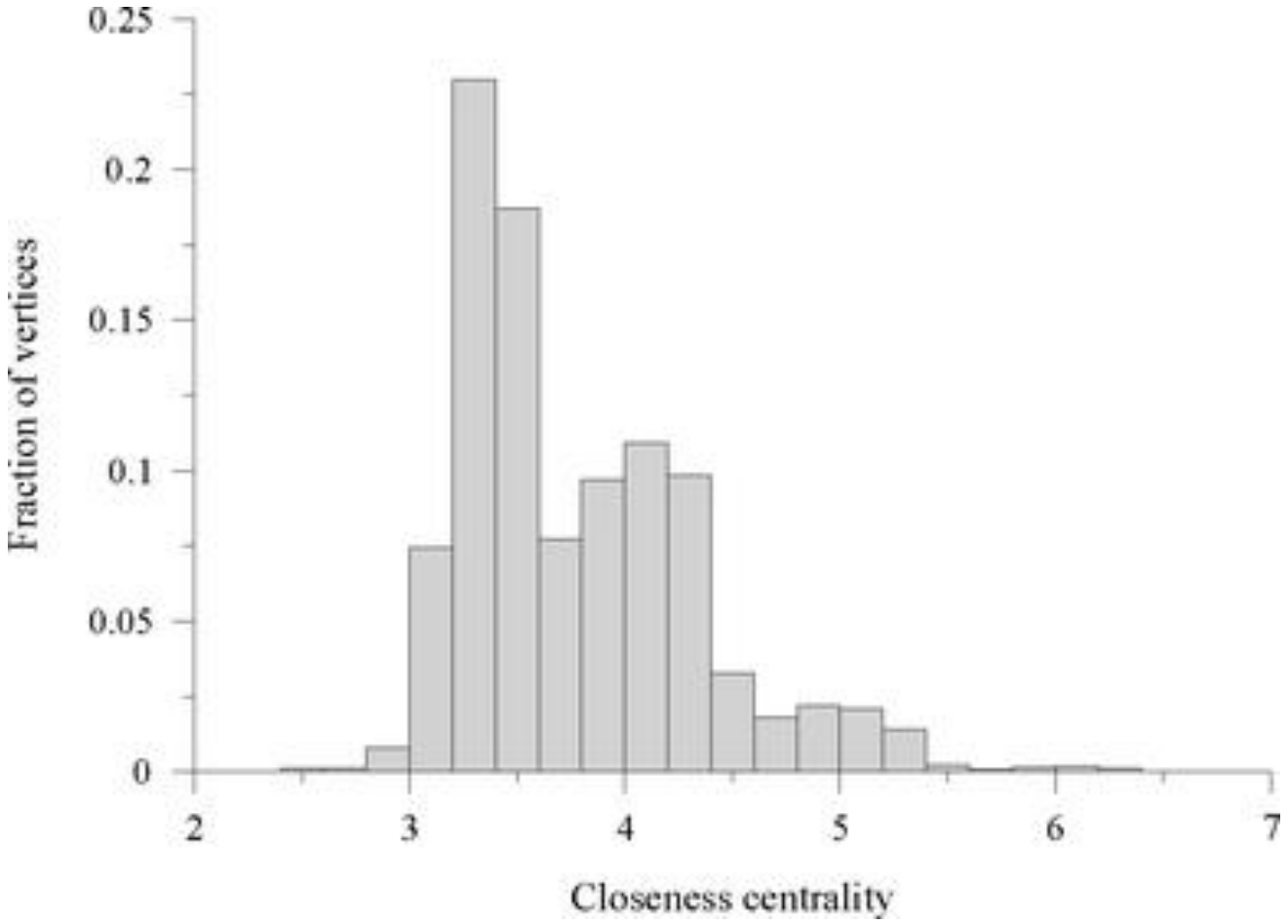
(c) Citation

Power-Law Cumulative Distribution Function for Two Centrality Measures of the Internet



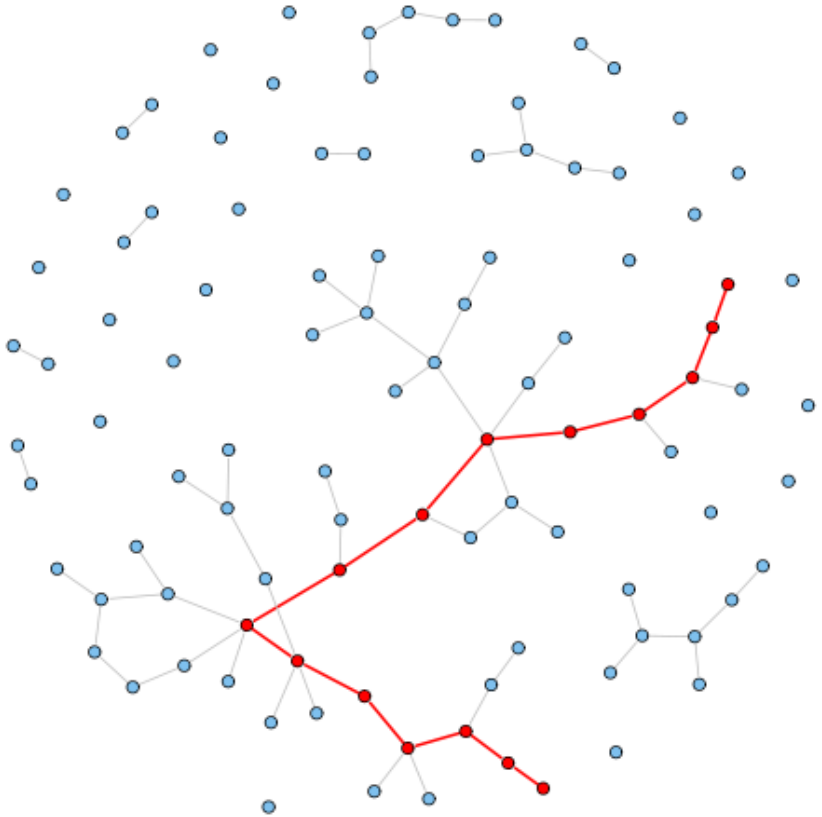
Not only degree centrality, but eigenvector and betweenness centralities exhibit a power law distribution for the internet network.

Histogram of the Closeness Centrality of the Internet: Not a Power Law



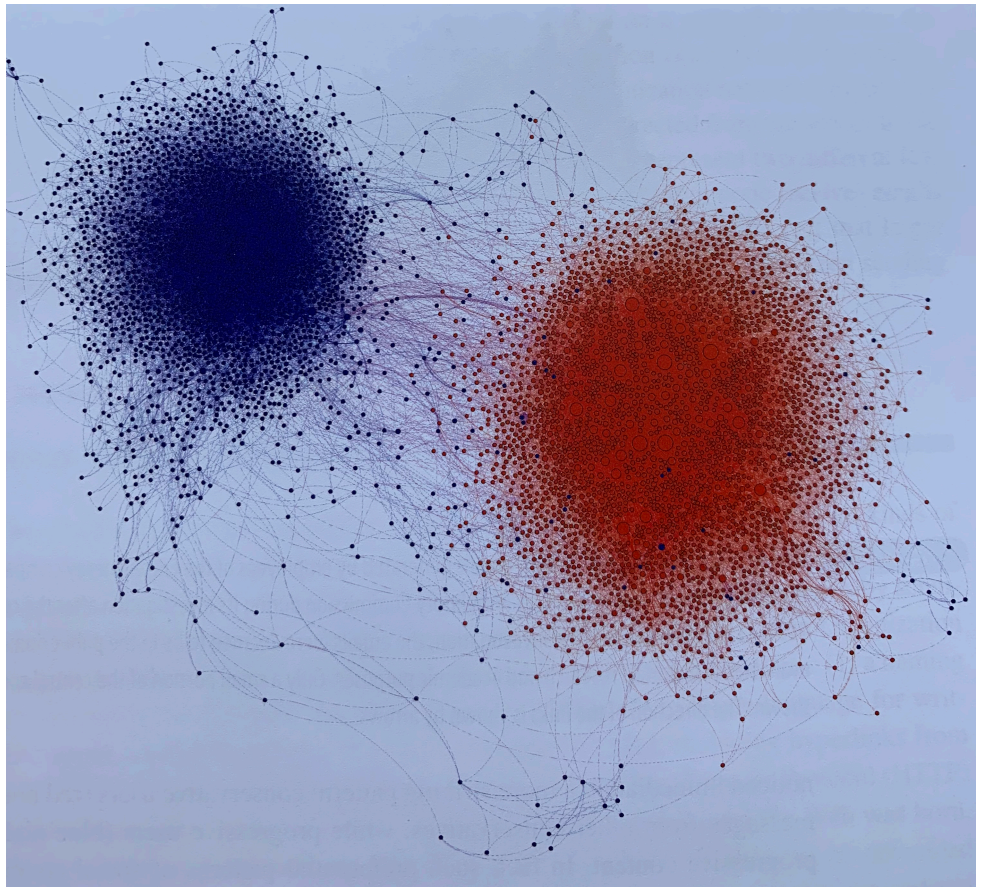
This centrality measure is limited by the diameter of the network, so it can't have the large right skew that occurs in power laws.

Assortativity and Short Geodesic Paths



Birds of a Feather Flock Together

Homophily refers to the tendency of people with common interests or geographic proximity to meet each other and become friends. This leads to friendship networks in which neighbors have similar characteristics. Such networks are said to exhibit assortativity.



Retweet network on Twitter among people sharing posts about US politics. Those with conservative content are in red, and those with progressive content are in blue. The echo chamber phenomenon is clear.

Menczer et al text

Assortativity Based on Degree

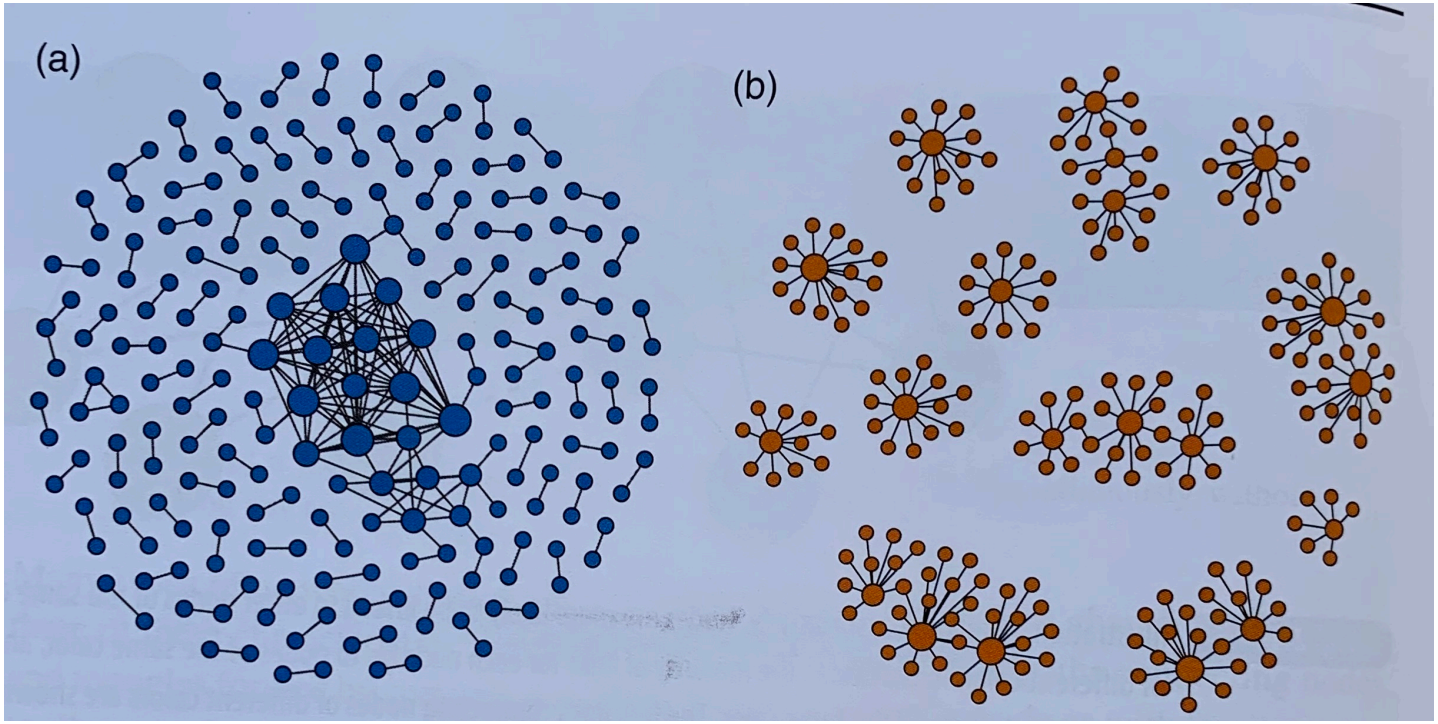
Do nodes with high degree tend to have neighbors that also have high degree?

If there is a positive correlation between the degree of a node and the degree of its neighbors, then the networks is assortative. They have a core-periphery structure.

If there is a negative correlation, then the network is disassortative.

Many real world networks are assortative.

Assortative and Disassortative Networks



Assortative network
(core-periphery structure)

Disassortative network

Measure of Assortativity

For each node *i* determine the average degree of its neighbors:

$$k_{nn}(i) = \frac{1}{k_i} \sum_j a_{ij} k_j$$

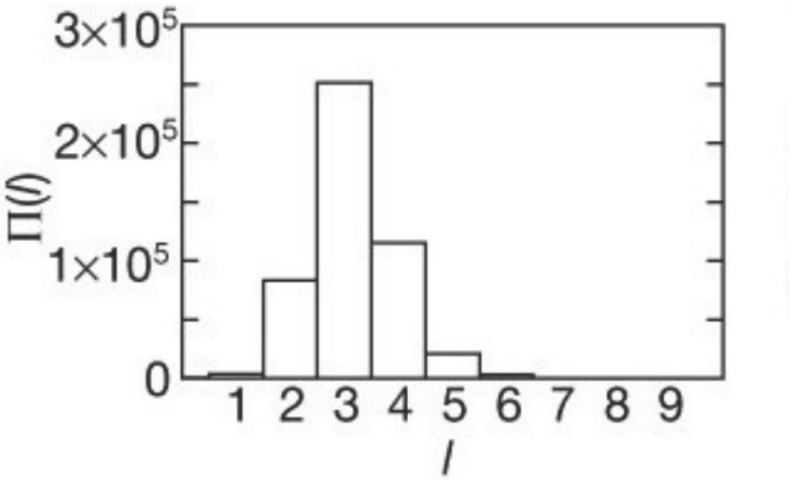
where $a_{ij} = 1$ if nodes *i* and *j* are neighbors, and 0 otherwise.

Define the *k*-nearest-neighbors function $\langle k_{nn}(k) \rangle$ for nodes of a given degree *k* as the average of $k_{nn}(i)$ across all nodes with degree *k*.

If $\langle k_{nn}(k) \rangle$ is an increasing function of *k*, then high-degree nodes tend to be connected to high-degree nodes and the network is assortative.
If $\langle k_{nn}(k) \rangle$ decreases with *k*, then the network is disassortative.

Network Diameter is Small in Real Networks, Even if the Network is Large

Short Geodesic Paths in Metabolic Networks



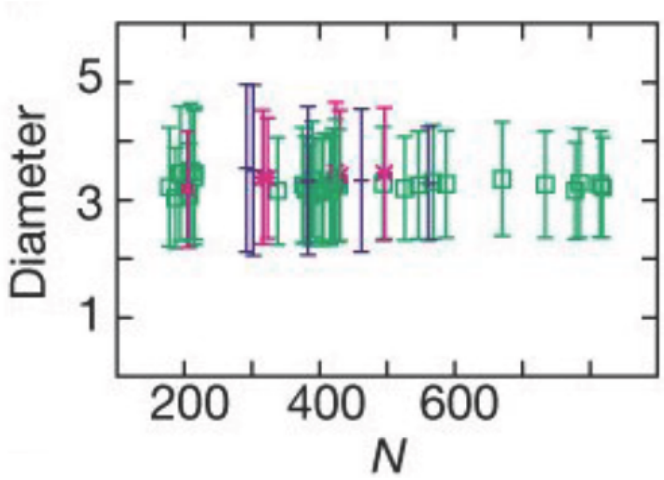
The histogram of the path lengths in the *E. coli* metabolic network

y-axis is number of geodesic paths,
x-axis is geodesic path length

Jeong et al., Nature, 407:651, 2000

Short Geodesic Paths in Metabolic Networks

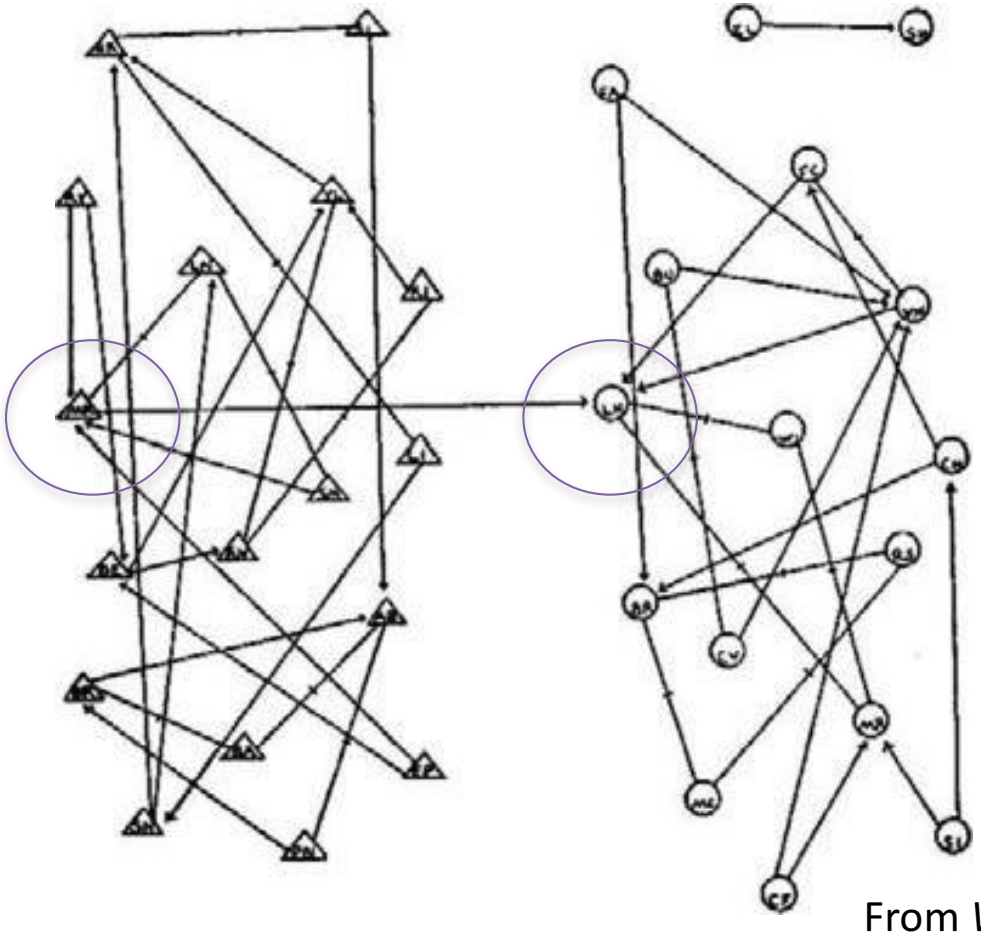
Here, network diameter means average geodesic path length for the network. The x-axis shows the number of nodes in the network.



The average path lengths for metabolic networks of 43 organisms with different complexity

Jeong et al., Nature, 407:651, 2000

Geodesic Paths Between Kids Varies, But is Usually Short



From *Who Shall Survive?*

Friendship network between a class of school children. Triangles=boys, Circles=girls.

Six Degrees of Separation

Social psychologist Stanley Milgram conducted a series of experiments in the 1960s to determine path lengths within friendship networks.

He randomly mailed 96 packages, or "passports" to recipients in Omaha, Nebraska. The goal was to get the passports to a lawyer friend of Milgram in Boston, the target.

Each recipient was asked to send the passport to the target (if they knew him), or, if not, to a friend who might know the target. They were told that he was a lawyer living in a eastern U.S. city. They would also write their name/address/occupation in the passport.

The path length from starting point to end point is defined as the number of names in the passport once it arrives to the target.

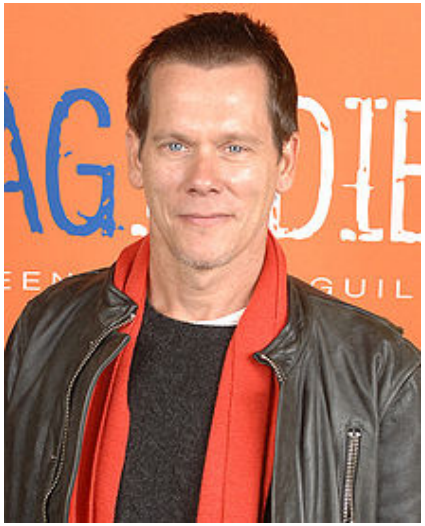
A total of 18 passports arrived to the target. What was the average path length? 5.9

Similar experiments with greater sample size have been performed since then. All give values close to 6. Thus, there is an average 6 degrees of separation between any two randomly chosen individuals in the world.

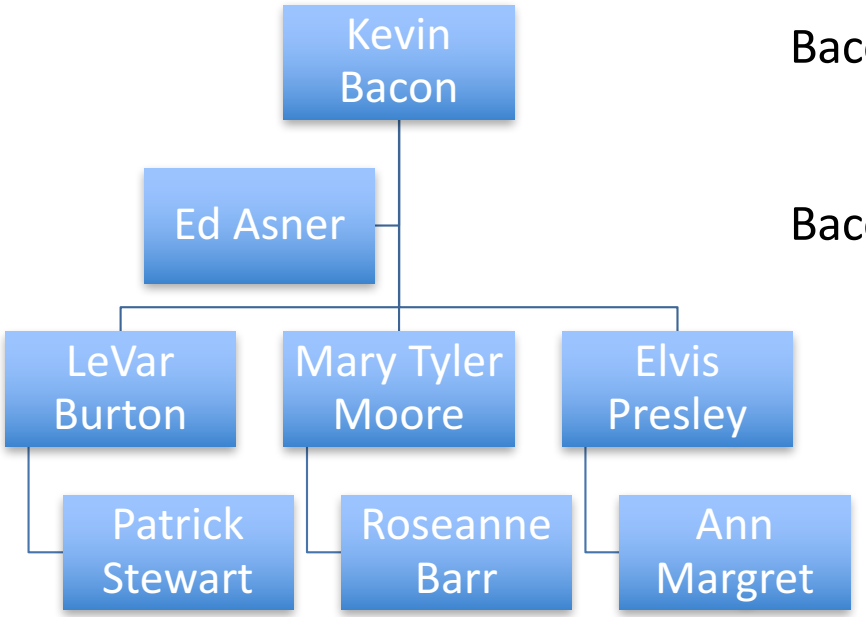
Six Degrees of Kevin Bacon

Bacon in 1994 interview: “I have worked with everyone in Hollywood or someone who has worked with them”

This motivated four college students at Albright College to create a network of those who have been in movies with Bacon, or those who have been in movies with them.



Six Degrees of Kevin Bacon



Bacon number = 0

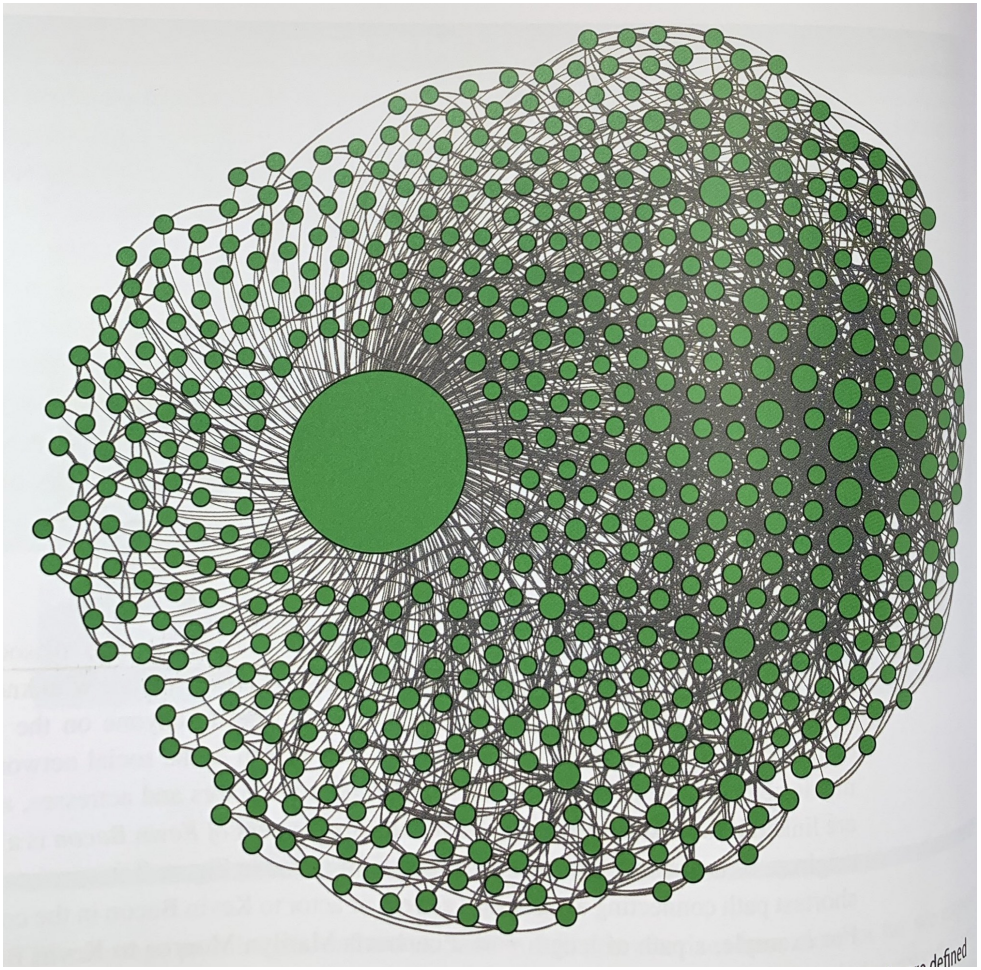
Bacon number = 1

Bacon number = 2

Bacon number = 3

http://oracleofbacon.org/

Mathematics Collaboration Network

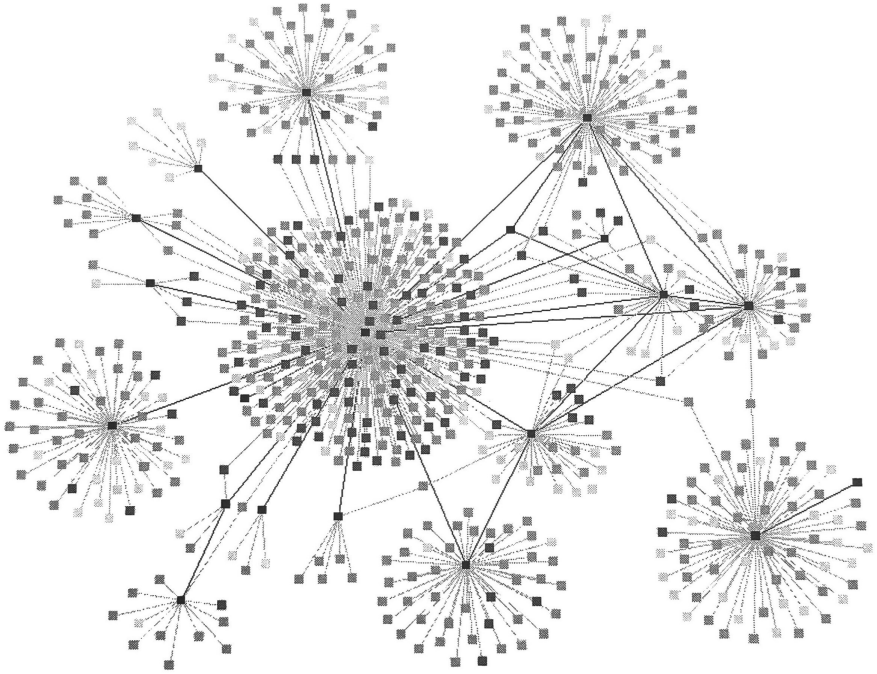


Menczer, Fortunato, Davis

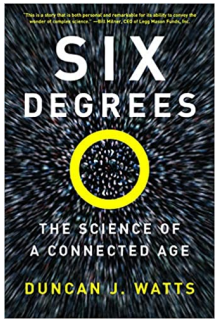
Ego network for mathematician Paul Erdos. Each node is a mathematician, each edge is a collaboration. Node size is based on degree.

http://www.ams.org/mathscinet/collaborationDistance.html

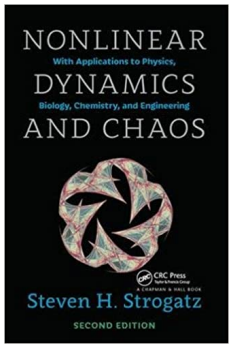
Why are Geodesic Path Lengths in Real-World Networks so Short?



This graph of the spread of tuberculosis has a network structure with lots of local connections and a few global connections. Such a network is called a small-world network, and is very common in real-world networks. These networks typically have low network diameter and a high clustering coefficient, which is relatively insensitive to the number of nodes in the network.

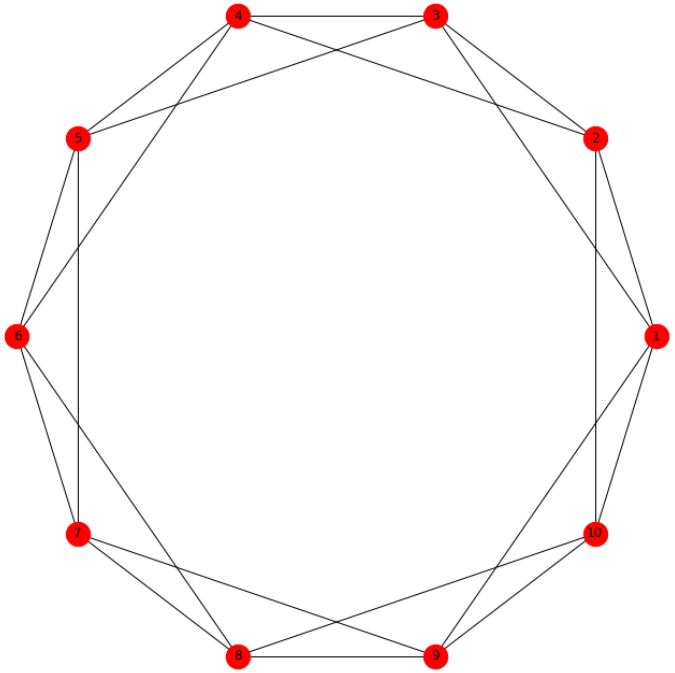


The Watts-Strogatz Model



This is an algorithm for creating a small-world network, starting from a ring network.

Each node has *k* edges, *k*/2 to nearest neighbors on each side. In this example, *k*=4.

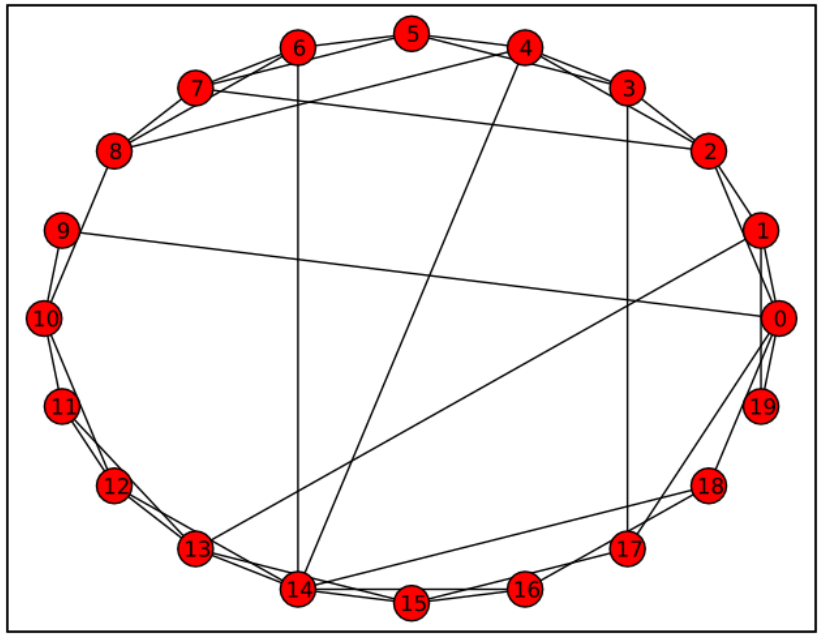


With this initial structure, geodesic paths are typically long, with mean of $l = \frac{n}{2k}$.

The clustering coefficient is high, with $C = \frac{3(k-2)}{4(k-1)}$, which tends to ¾ for large *k*.

The Watts-Strogatz Model

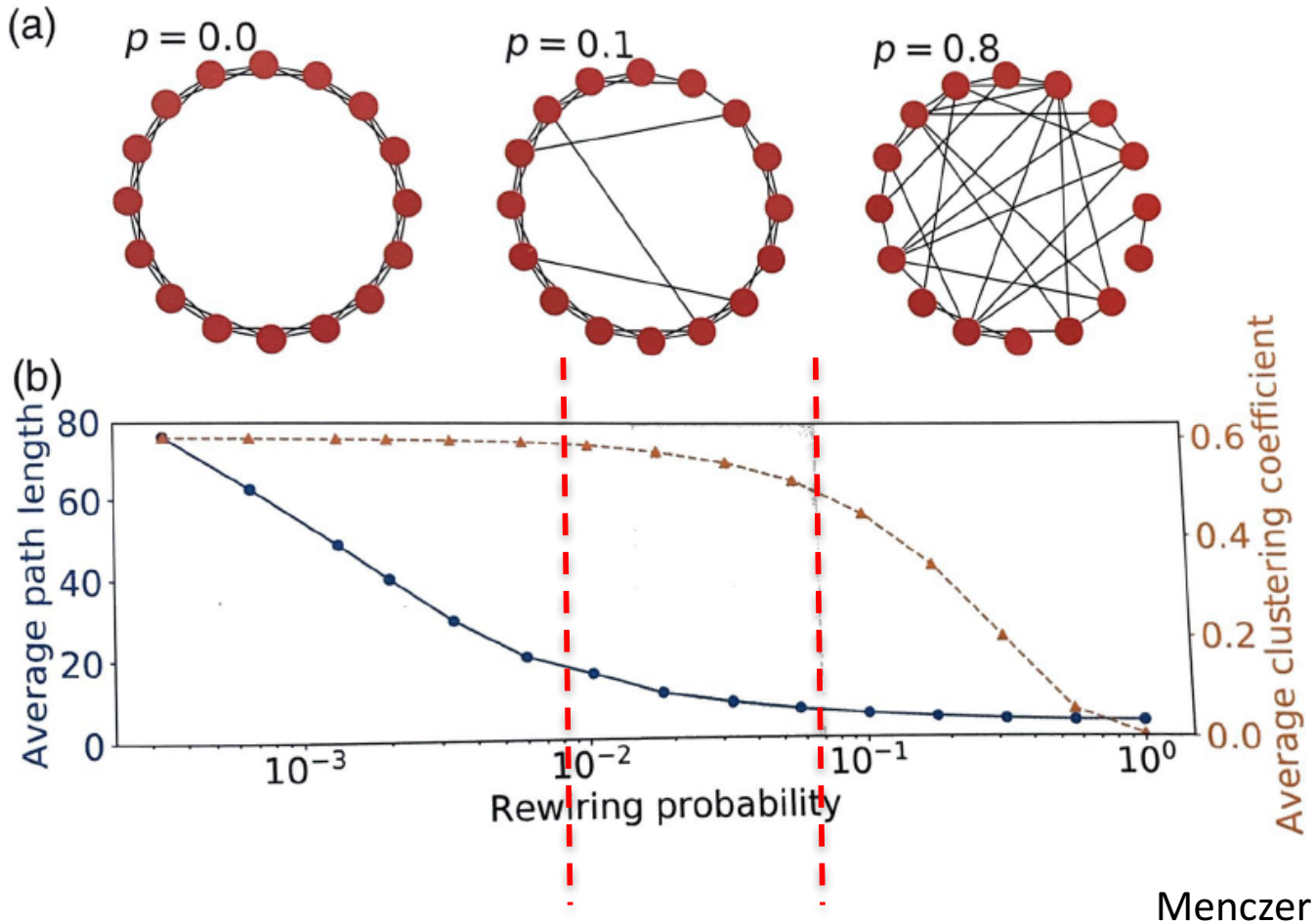
Next, move along each node. For each of its *k* edges, rewire it from the neighbor to a randomly chosen node with probability p.



The network now has a lot of local connections and a few global connections, making it a small-world network.

The Watts-Strogatz Model

With $p = 0$ the graph is a regular grid. With $p = 1$ it is a random graph.
In some intermediate range of p values it has the high clustering coefficient of a regular grid, but the low geodesic path length of a random graph.
In this range, the graph has the small world property.



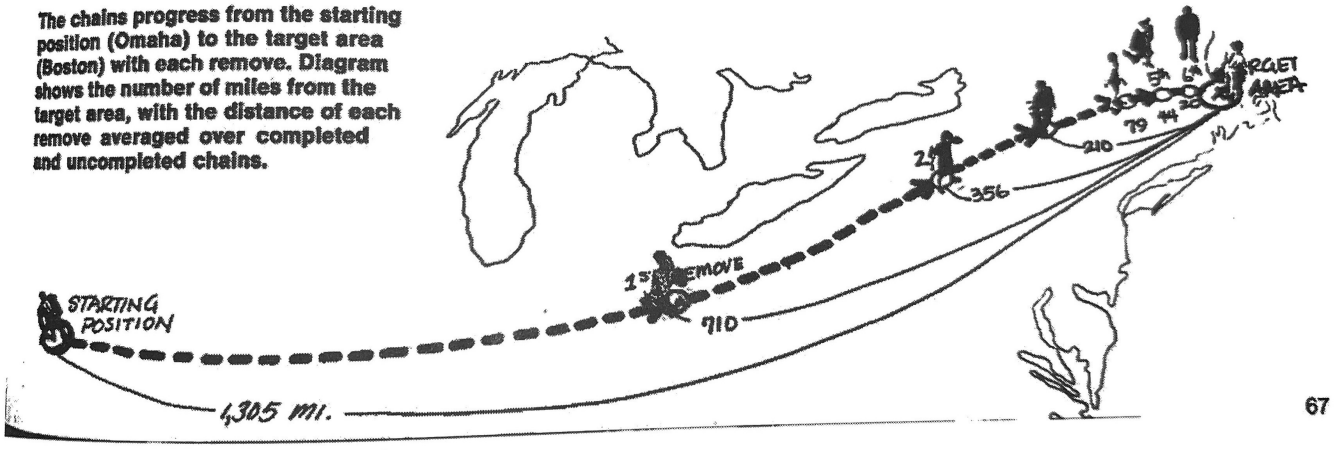
Menczer et al., 2020

Efficient Decentralized Searches

Centralized network

Decentralized network

Milgram Experiment Really Showed Two Things

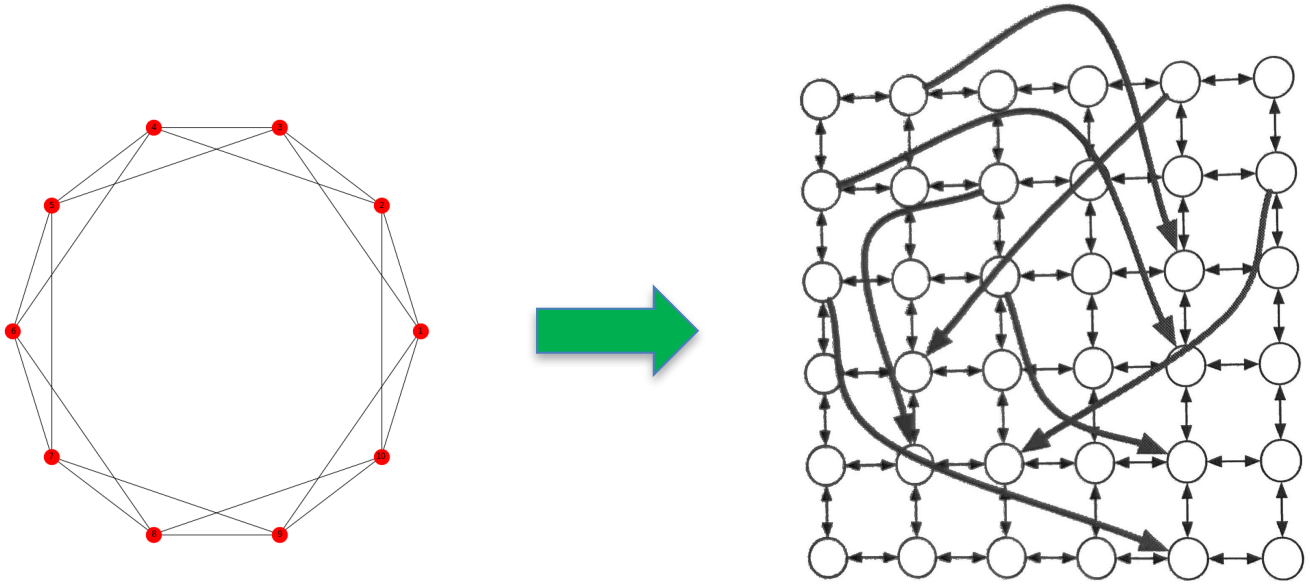


First, it showed that there are some very short paths in social networks. This is the *small world phenomenon*.

Second, it showed that **people somehow find these short paths**, even though all they know about the connections of the network is their own neighborhood. Since the social network is decentralized, this is called a decentralized search. Why were the participants so good at finding short paths to the target?

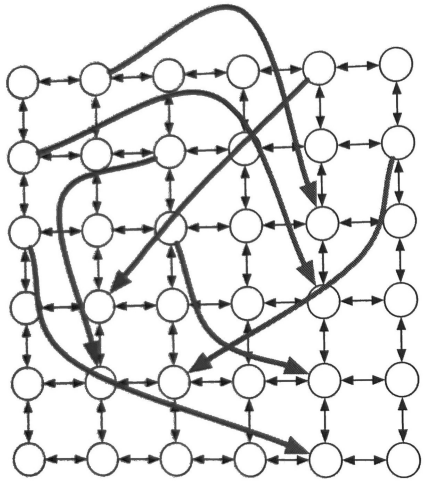
Try It With a Watts-Strogatz Model

In this implementation, place the nodes on a rectangular grid and randomly add M weak (long-range) connections.



If you start at some node *v* with target *w*, and randomly choose an edge out of *v* to some neighbor, repeating at that neighbor, then the average path length taken is much greater than the geodesic path length from *v* to *w*. That is, the decentralized search over this small world network did not find geodesic paths to the target.

The Weak Links Are Too Random



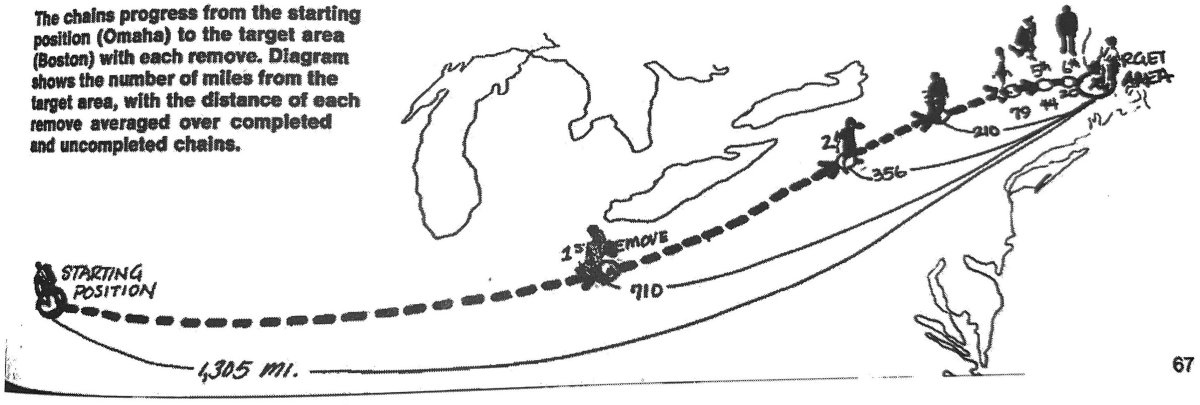
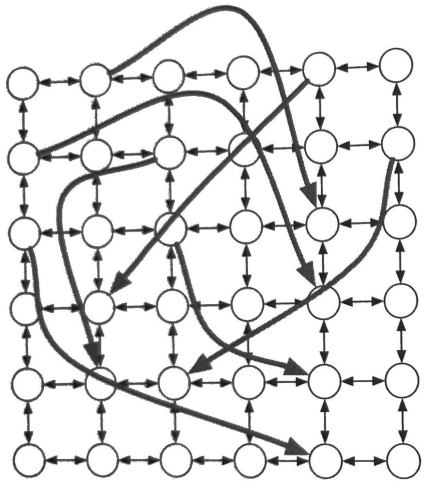


Figure from Milgram's 1967 article in *Psychology Today*

The people receiving letters knew that the target lived in Boston and was a lawyer. This information biased their decision on who to send the letter to. Each step took the letter closer to the target. This would not be the case in the standard Watts-Strogatz network.

It is not enough to have a network model with weak ties over very long ranges. Also need to span the intermediate ranges.

Solution: Bias the Selection of Weak Ties



There are a lot more nodes far from a typical node *v* than there are close to *v*. So picking weak ties randomly favors the distant nodes.

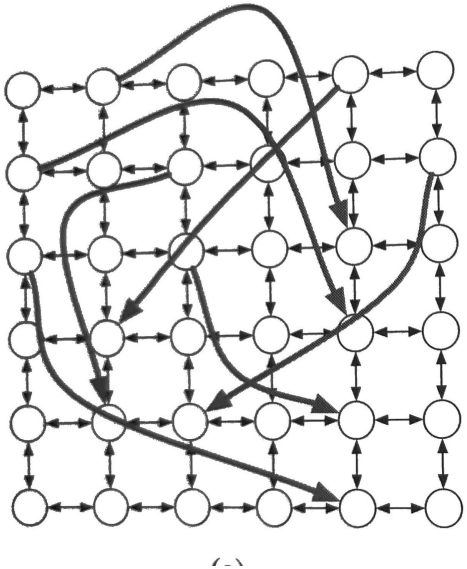
Let *d(v,w)* be the distance between nodes *v* and *w*. This distance is the number of edges taken on the geodesic path between the nodes only following the local edges.

Then determine weak connection placement such that the probability of connecting *v* to *w* is proportional to $\frac{1}{d(v,w)^q}$ where *q* is the clustering exponent. Nodes at greater distance will have lower probability of being connected if $q > 0$.

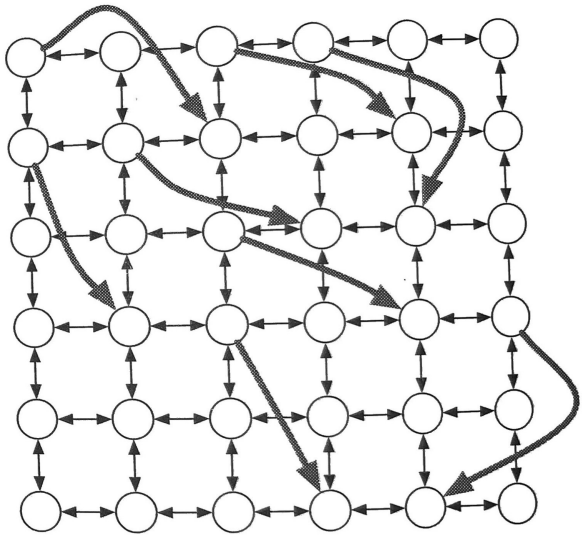
Which value of the clustering exponent gives shortest transit time? That is, the shortest number of steps from a starting point to a randomly-determined target?

Solution: Bias the Selection of Weak Ties

q small

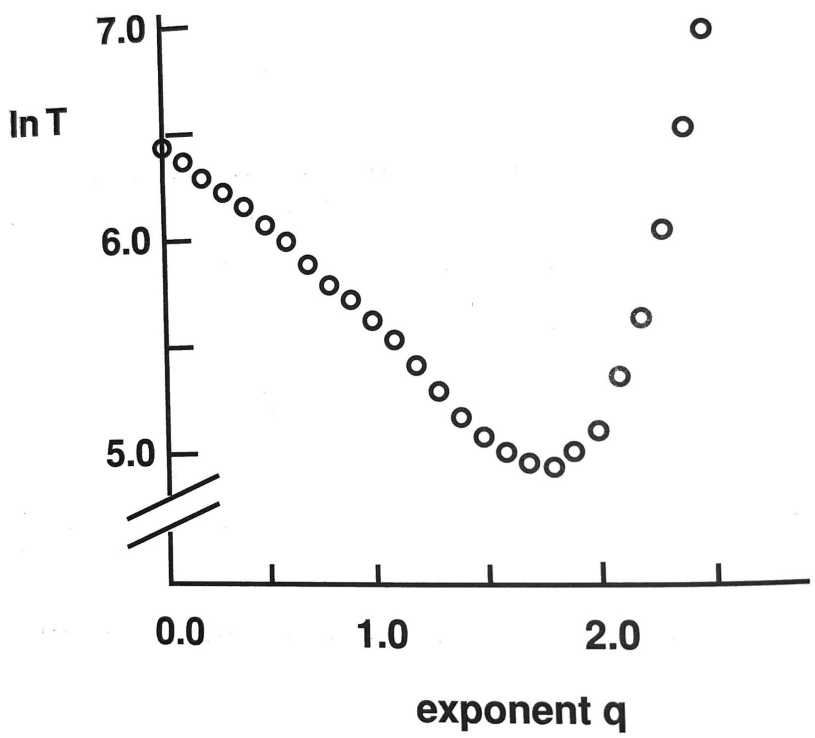


q large



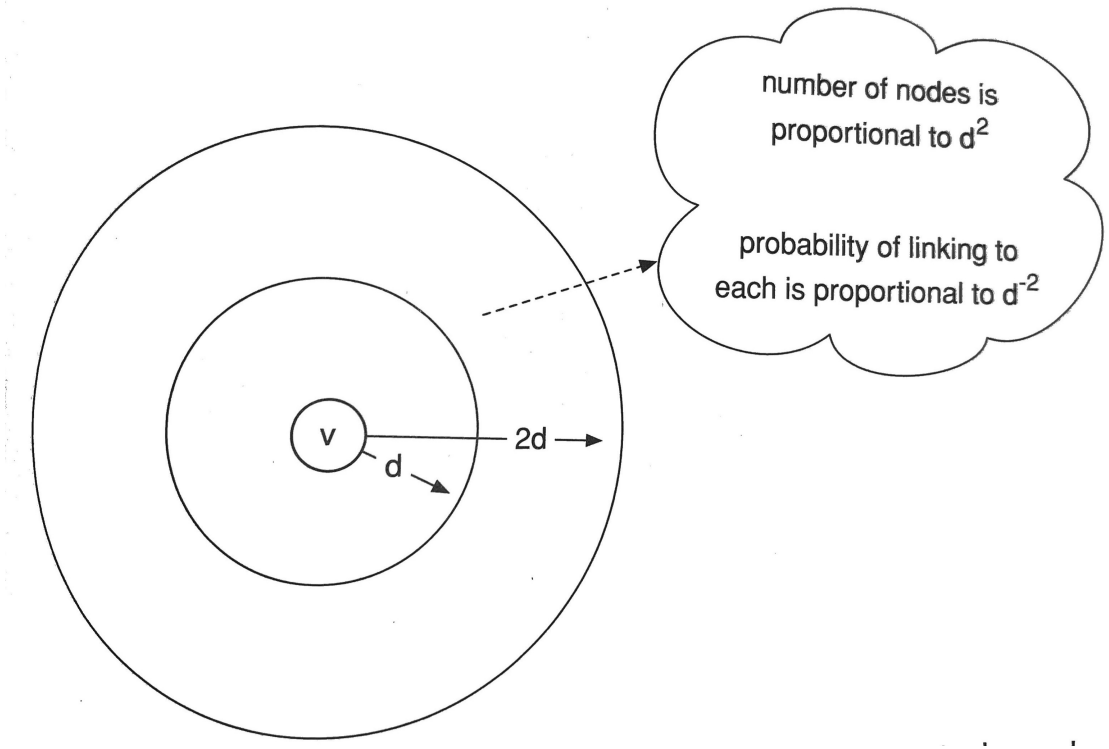
Solution: Bias the Selection of Weak Ties

Results from 400 million nodes, each point is the average of 1,000 runs.
T is the transit time between randomly selected nodes.



The best value of the clustering exponent is between 1.5 and 2. In the limit of infinitely large networks, the optimal clustering exponent is $q = 2$.

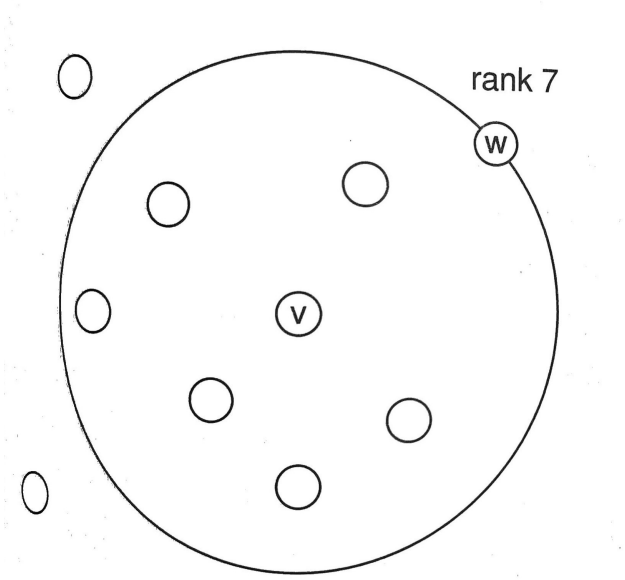
Why is a Clustering Exponent of 2 Optimal?



Consider neighborhoods of radius *d* and 2*d* about node *v*. The number of nodes in the ring is proportional to d^2. If random connections from *v* are proportional to $\frac{1}{d^2}$, then the probability that there is a random weak connection from *v* to a node in the periphery is the same as a connection from *v* to a node in the core. This is in spite of the fact that there are more nodes in the periphery. Thus, the connection is uniform across spatial scales.

How Does This Generalize to Non-uniform Distributions of Nodes?

A uniform grid with additional weak connections is a simple model to work with, but not realistic for most real-world situations. How can the results be generalized to non-uniform networks?



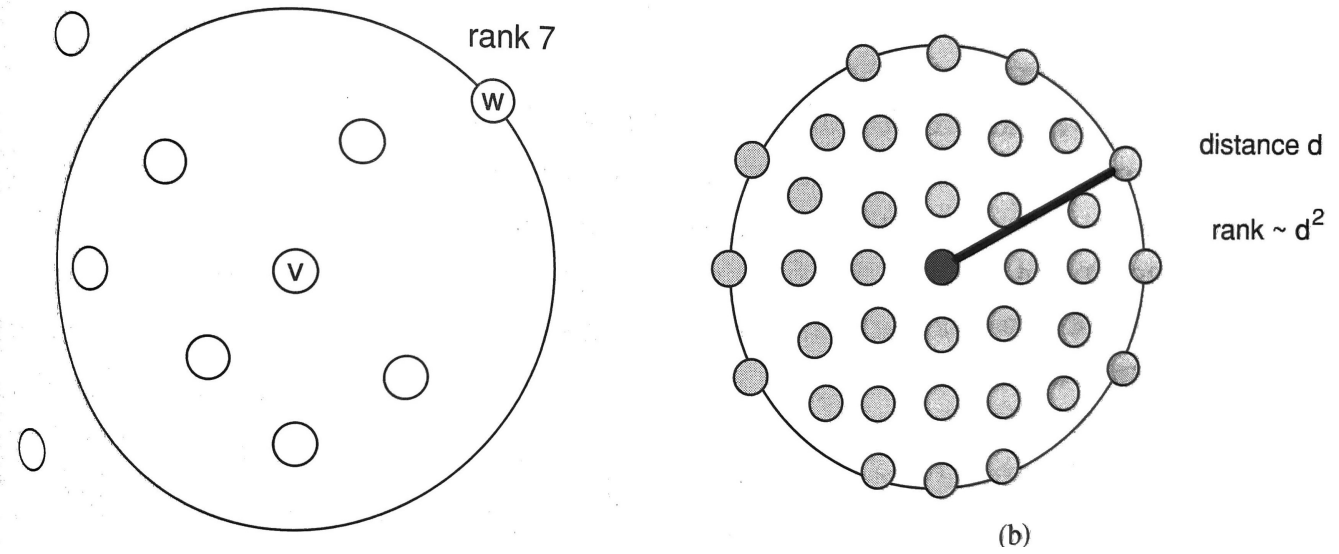
Define rank(*w*) as the number of closer nodes to *v*, including *v* itself.

How Does This Generalize to Non-uniform Distributions of Nodes?

Then select weak links in the network with probability proportional to $\frac{1}{\text{rank}(w)^p}$.

This is called rank-based friendship.

What is the optimal value of *p* to ensure optimal transit times in the network?



On a uniform grid, a node at distance *d* from a central node *v* would have rank proportional to d^2. So the optimal value of *p* would be $p = 1$.

Main Result

To construct a network that is efficiently searchable, create a link to each node with probability that is inversely proportional to the number of closer nodes.

What Does “Closer Nodes” Mean?

In terms of the Milgram experiment, “close” has a geographical meaning. However, in most contexts, geographical location is not the main factor in determining social contacts. What is?

A social focus is defined as any type of community, occupational pursuit, shared interest, or activity that organizes social life around it.

Some examples relevant to us:

We all live in Florida

We are all students or professors at FSU

We are all interested in mathematics

We all participate in this course

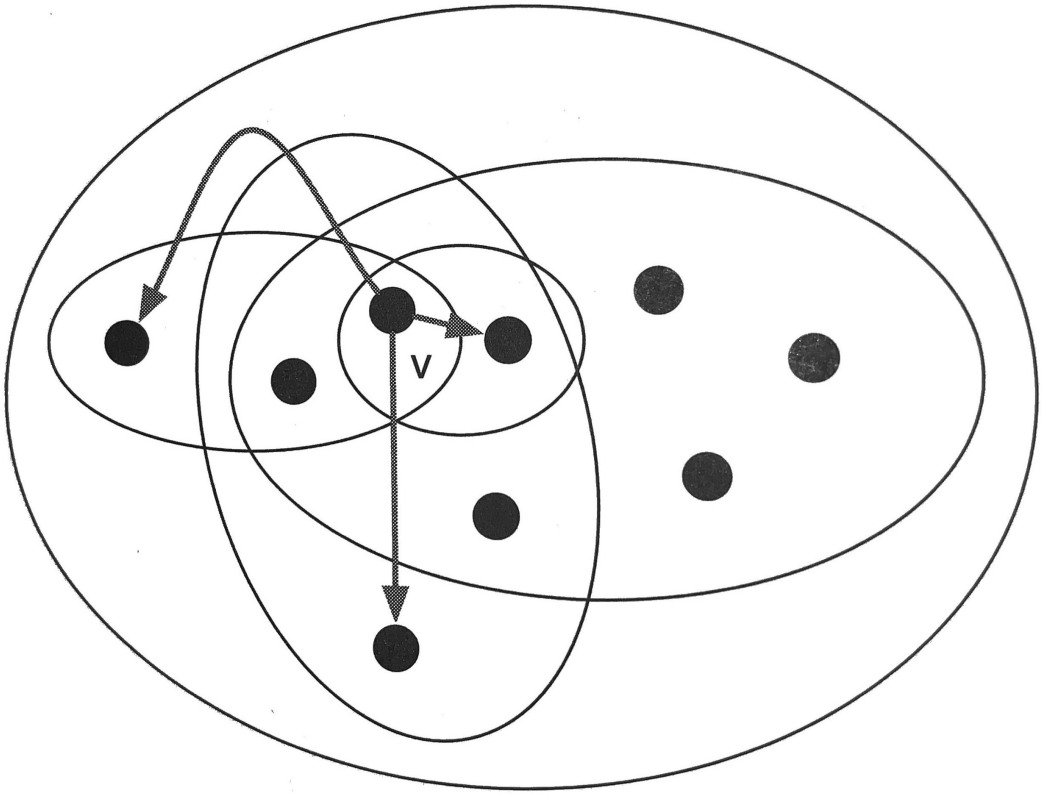
The Concept of Social Distance

Participation in this course is the most likely to result in the formation of social connections. This is largely because the number of participants in the course is very small, making it likely that the participants will get to know one another.

This example motivates the definition of social distance between two nodes *v* and *w* as the size of the smallest social focus that contains both nodes.

Social Focus Overlap and Social Distance

What are the social distances between *v* and the three nodes indicated?

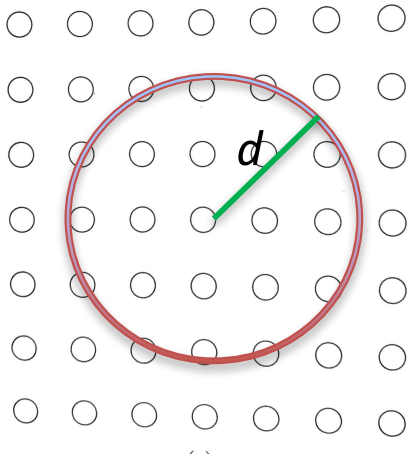


The social distance between *v* and each of the three nodes indicated is 2, 3, and 5.

Optimal Searching Using Social Distance

Let *sd(v,w)* be the social distance between nodes *v* and *w*. It has been shown that if links are generated in the network with probability proportional to $\frac{1}{sd(v,w)}$, then the network supports efficient decentralized search.

Why $\frac{1}{sd(v,w)}$ and not $\frac{1}{sd(v,w)^2}$?



Return to a uniform grid. The number of nodes in a circle of radius *d* is proportional to d^2. That is, $sd(v,w) \propto d(v,w)^2$. Since optimal edge placement occurs when placement probability is proportional to $\frac{1}{d(v,w)^2}$, it is therefore optimal with placement probability proportional to $\frac{1}{sd(v,w)}$.

The End