Web Mining ed Analisi delle Reti Sociali

Modelli di generazione delle reti

Dino Pedreschi
Dipartimento di Informatica
Università di Pisa
www.di.unipi.it/~pedre

Social Network Analysis

- Social Network Introduction
- Statistics and Probability Theory
- Models of Social Network Generation
- Mining on Social Network
- Summary

Some Models of Network Generation

- *Random graphs (Erdös-Rényi models):*
  - gives few components and small diameter
  - does not give high clustering and heavy-tailed degree distributions
  - is the mathematically most well-studied and understood model
- *Watts-Strogatz models:*
  - give few components, small diameter and high clustering
  - does not give heavy-tailed degree distributions
- *Scale-free Networks:*
  - gives few components, small diameter and heavy-tailed distribution
  - does not give high clustering
- *Hierarchical networks:*
  - few components, small diameter, high clustering, heavy-tailed
- *Affiliation networks:*
  - models group-actor formation

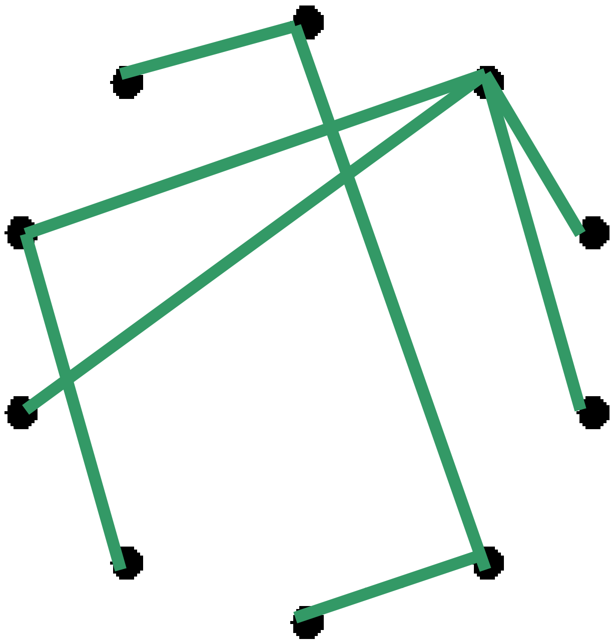
Models of Social Network Generation

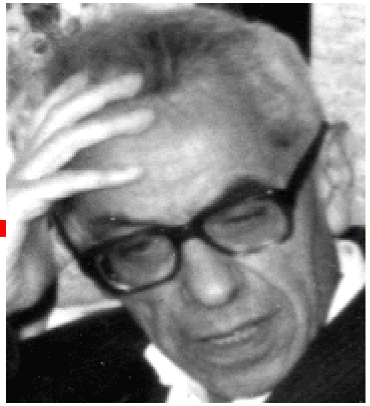
- Random Graphs (Erdös-Rényi models)
- Watts-Strogatz models
- Scale-free Networks

The Erdös-Rényi (ER) Model (Random Graphs)

- 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 ~ 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 → infinity
  - 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) ~ pN
  - Sharply concentrated; *not* heavy-tailed
- Especially easy to *generate* NWs from G(N,p)

Erdös-Rényi Model (1960)



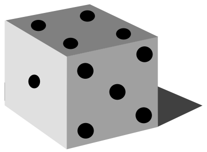


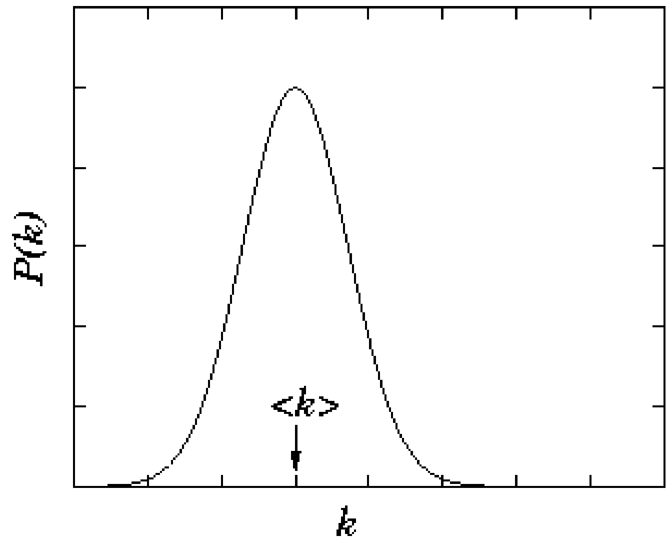
Connect with probability p

Pál Erdös
(1913-1996)

p=1/6
N=10
⟨k⟩~1.5

Poisson distribution





- **Democratic**
- **Random**

A Closely Related Model

- For any fixed m <= 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) ~ 2m/N^2
  - this intuition can be made precise, and is correct
  - if m = cN then p = 2c/(N-1) ~ 2c/N
  - mathematically trickier than G(N,p)

Another Closely Related Model

- *Graph process* model:
  - start with N vertices and no edges
  - at each time step, add a *new* edge
  - choose new edge randomly from among all missing edges
- Allows study of the *evolution* or *emergence* of properties:
  - as the number of edges m grows in relation to N
  - equivalently, as p is increased
- For all of these models:
  - high probability ←→ "almost all" large graphs of a given density

Evolution of a Random Network

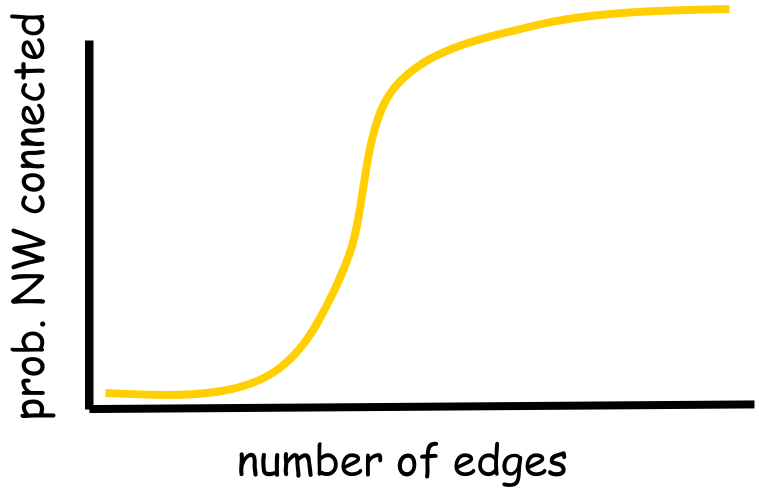
- We have a large number n of vertices
- We start randomly adding edges one at a time
- At what time t will the network:
  - *have at least one "large" connected component?*
  - *have a single connected component?*
  - *have "small" diameter?*
  - *have a "large" clique?*
  - *have a "large" chromatic number?*
- How gradually or suddenly do these properties appear?

Recap

- Model G(N,p):
  - select each of the possible edges independently with prob. p
  - *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 (*not* heavy-tailed)
- Model G(N,m):
  - select *exactly m* of the N(N-1)/2 edges to appear
  - all sets of m edges equally likely
- Graph process model:
  - starting with no edges, just keep adding one edge at a time
  - always choose next edge randomly from among all missing edges
- Threshold or tipping for (say) connectivity:
  - fewer than m(N) edges → graph almost certainly *not* connected
  - more than m(N) edges → graph almost certainly *is* connected
  - made formal by examining limit as N → infinity

Combining and Formalizing Familiar Ideas

- Explaining *universal behavior* through statistical models
  - our models will always generate *many* networks
  - *almost all* of them will share certain properties (universals)
- Explaining *tipping* through incremental growth
  - we gradually add edges, or gradually increase edge probability p
  - many properties will emerge *very suddenly* during this process



Monotone Network Properties

- Often interested in *monotone* graph properties:
  - let G have the property
  - *add edges* to G to obtain G'
  - then G' *must* have the property also
- Examples:
  - G is connected
  - G has diameter <= d (*not* exactly d)
  - G has a clique of size >= k (*not* exactly k)
  - G has chromatic number >= c (*not* exactly c)
  - G has a matching of size >= m
  - d, k, c, m 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?

Formalizing Tipping: Thresholds for Monotone Properties

- 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 → G has the property
  - P(G) = 0 → G does not have the property
- Let m(N) be some function of NW 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 → infinity
  - if r(N) → 0: probability that P(G) = 1 in G(N,m'(N)): → 0
  - if r(N) → infinity: probability that P(G) = 1 in G(N,m'(N)): → 1
- A *purely structural* definition of tipping
  - tipping results from incremental increase in *connectivity*

So Which Properties Tip?

- Just about *all* of them!
- The following properties all have threshold functions:
  - having a “giant component”
  - being connected
  - having a perfect matching (N even)
  - having “small” diameter
- With remarkable consistency (N = 50):
  - giant component ~ 40 edges, connected ~ 100, small diameter ~ 180

Ever More Precise...

- Connected component of size > N/2:
  - threshold function is m(N) = N/2 (or p ~ 1/N)
  - note: full connectivity *impossible*
- Fully connected:
  - threshold function is m(N) = (N/2)log(N) (or p ~ log(N)/N)
  - NW remains *extremely sparse:* only ~ log(N) edges per vertex
- Small diameter:
  - threshold is m(N) ~ N^(3/2) for *diameter 2* (or p ~ 2/sqrt(N))
  - fraction of possible edges still ~ 2/sqrt(N) → 0
  - generate *very* small worlds

Other Tipping Points?

- Perfect matchings
  - consider only even N
  - threshold function: m(N) = (N/2)log(N) (or p ~ log(N)/N)
  - same as for connectivity!
- Cliques
  - k-clique threshold is m(N) = (1/2)N^(2 – 2/(k-1)) (p ~ 1/N^(2/k-1))
  - edges appear immediately; triangles at N/2; etc.
- Coloring
  - k colors required just as k-cliques appear

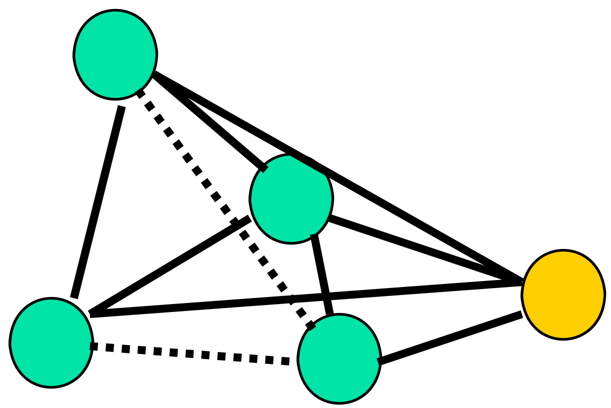
Erdos-Renyi Summary

- 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!

The Clustering Coefficient of a Network

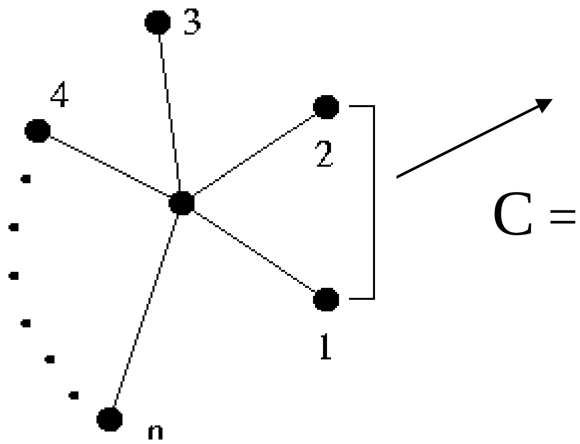
- Let nbr(u) denote the set of *neighbors* of u in a graph
  - all vertices v such that the edge (u,v) is in the graph
- The clustering coefficient of u:
  - let k = |nbr(u)| (i.e., number of neighbors of u)
  - choose(k,2): max possible # of edges between vertices in nbr(u)
  - c(u) = (*actual* # of edges between vertices in nbr(u))/choose(k,2)
  - 0 <= c(u) <= 1; measure of *cliquishness* of u's neighborhood
- Clustering coefficient of a graph:
  - average of c(u) over all vertices u

k = 4
choose(k,2) = 6
c(u) = 4/6 = 0.666…



The Clustering Coefficient of a Network

Clustering: My friends will likely know each other!



Probability to be connected C » p

$$C = \frac{\text{\# of links between 1,2,…n neighbors}}{n(n-1)/2}$$

Networks are clustered [large C(p)] but have a small characteristic path length [small L(p)].

Network	C	C_{rand}	L	N
WWW	0.1078	0.00023	3.1	153127
Internet	0.18-0.3	0.001	3.7-3.76	3015-6209
Actor	0.79	0.00027	3.65	225226
Coauthorship	0.43	0.00018	5.9	52909
Metabolic	0.32	0.026	2.9	282
Foodweb	0.22	0.06	2.43	134
C. elegance	0.28	0.05	2.65	282

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 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

Models of Social Network Generation

- Random Graphs (Erdös-Rényi models)
- Watts-Strogatz models
- Scale-free Networks

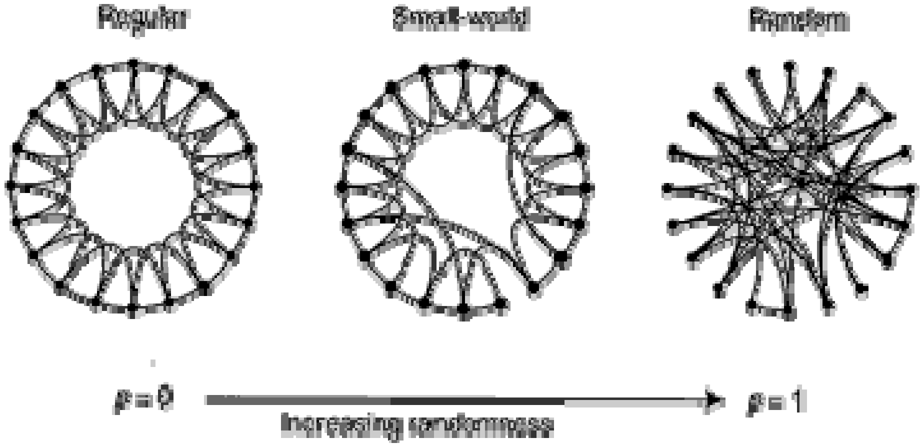
Caveman and Solaria

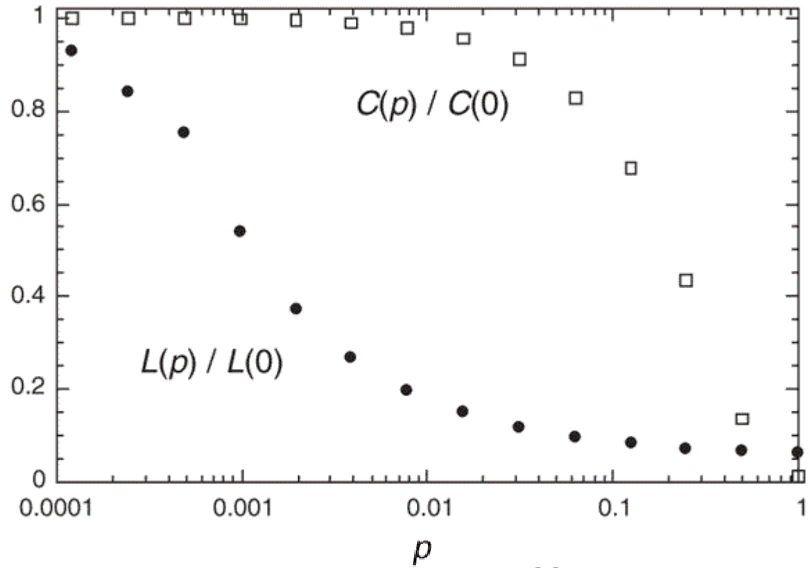
- Erdos-Renyi:
  - sharing a common neighbor makes two vertices *no more likely* to be directly connected than two very "distant" vertices
  - every edge appears entirely *independently* of existing structure
- But in many settings, the *opposite* is true:
  - you tend to meet new friends through your 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

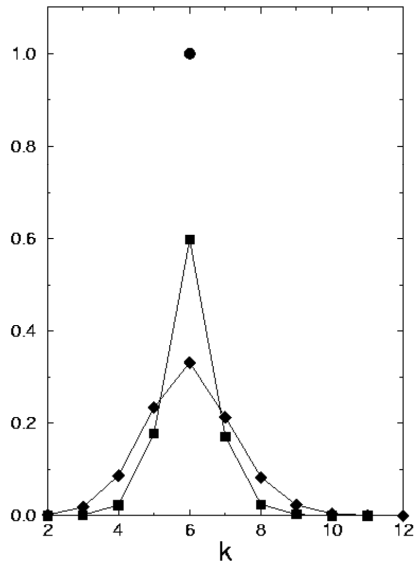
Making it (Somewhat) Precise: the α-model

- The α-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
  - α: 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) >= k, R(u,v) = 1 (share too many friends *not* to connect)
  - if m(u,v) = 0, R(u,v) = p (no mutual friends → 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: α = infinity is "like" Erdos-Renyi (but not exactly)

Watts-Strogatz Model







C(p) : clustering coeff.

L(p) : average path length

(Watts and Strogatz, Nature **393,** 440 (1998))

Small Worlds and Occam's Razor

- For small α, 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 → small diameter
  - α-model → 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 C. elegans nervous system
- For each of these networks, he:
  - computes its size, diameter, and clustering coefficient
  - compares diameter and clustering to *best* Erdos-Renyi approx.
  - shows that the *best* α-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, α is better than p

Case 1: Kevin Bacon Graph

- Vertices: actors and actresses
- Edge between u and v if they appeared in a film together

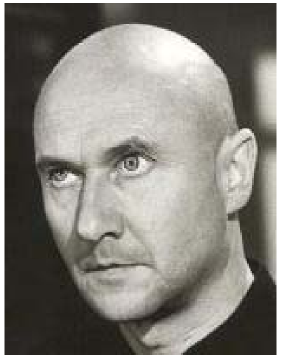
Kevin Bacon
No. of movies : 46 No. of actors : 1811 Average separation: 2.79

Is Kevin Bacon the most connected actor?

NO!

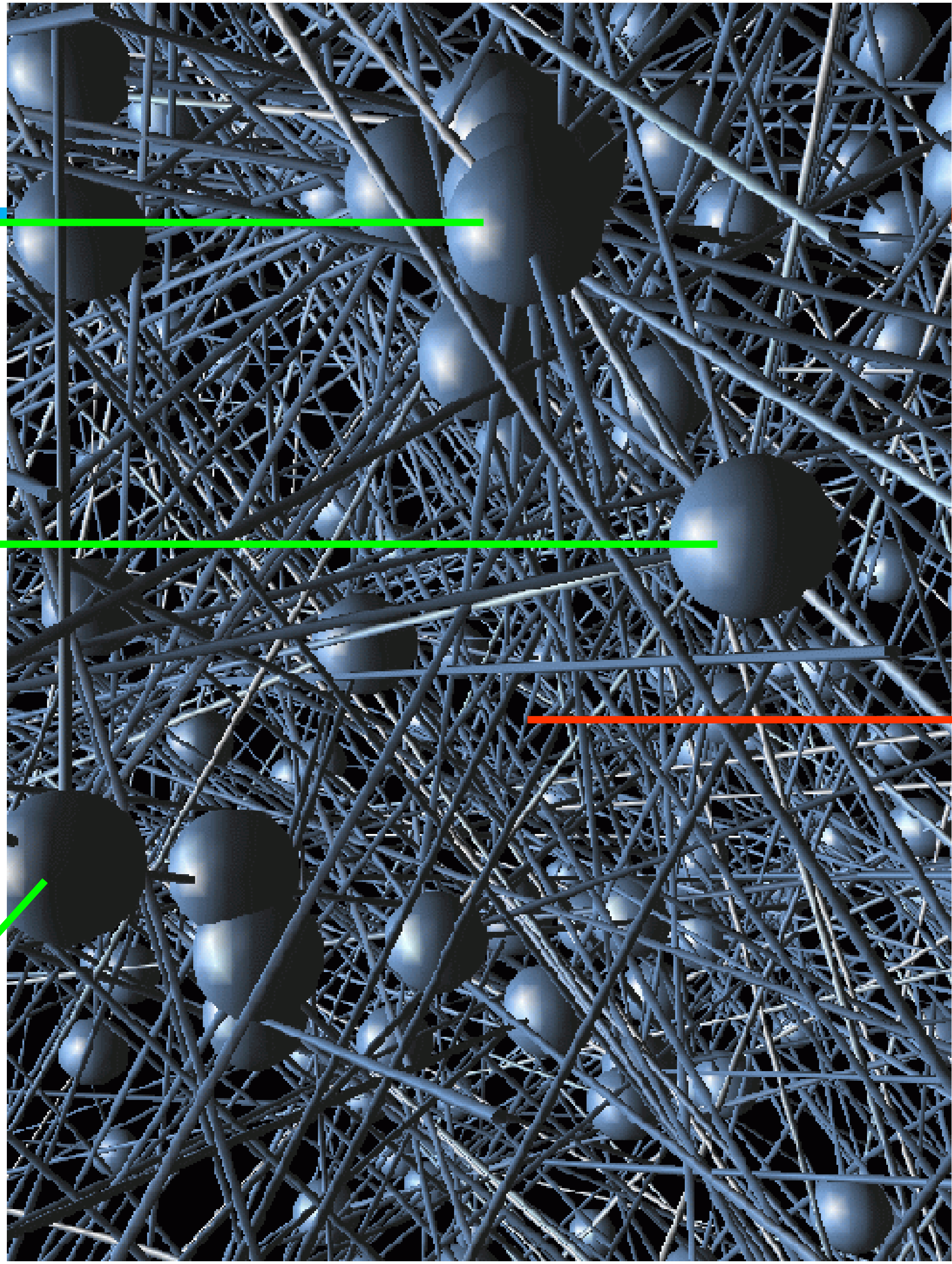
Rank	Name	Average distance	# of movies	# of links
1	Rod Steiger	2.537527	112	2562
2	Donald Pleasence	2.542376	180	2874
3	Martin Sheen	2.551210	136	3501
4	Christopher Lee	2.552497	201	2993
5	Robert Mitchum	2.557181	136	2905
6	Charlton Heston	2.566284	104	2552
7	Eddie Albert	2.567036	112	3333
8	Robert Vaughn	2.570193	126	2761
9	Donald Sutherland	2.577880	107	2865
10	John Gielgud	2.578980	122	2942
11	Anthony Quinn	2.579750	146	2978
12	James Earl Jones	2.584440	112	3787
…				
876	**Kevin Bacon**	**2.786981**	**46**	**1811**
…				

#1 Rod Steiger



#2 Donald Pleasence

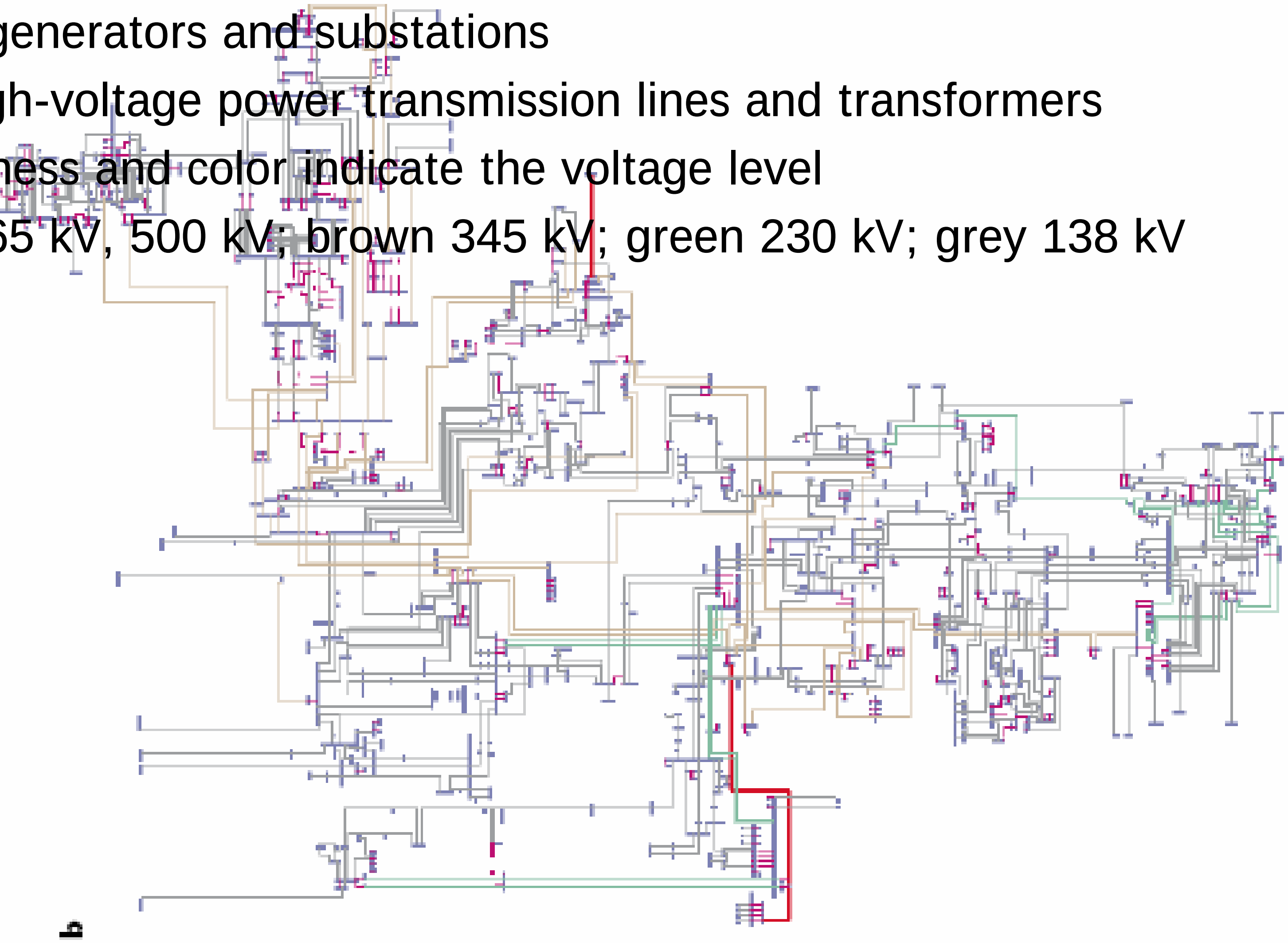
#3 Martin Sheen



#876 Kevin Bacon

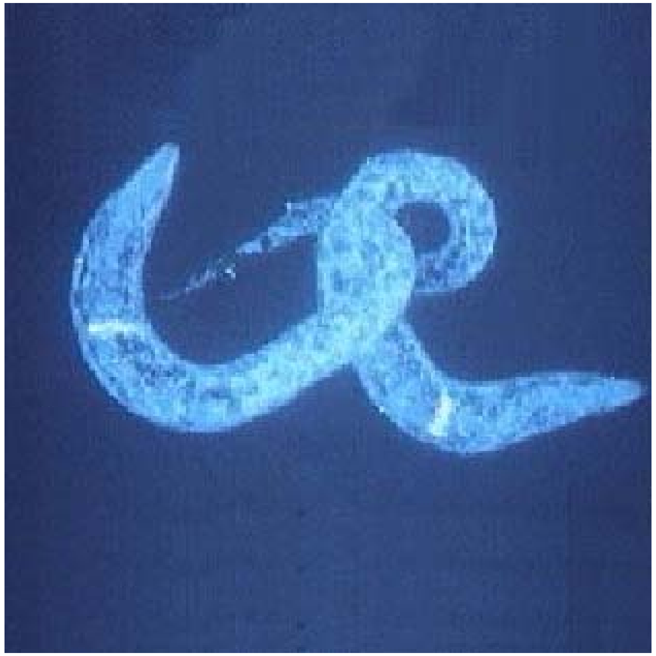
Case 2: New York State Power Grid

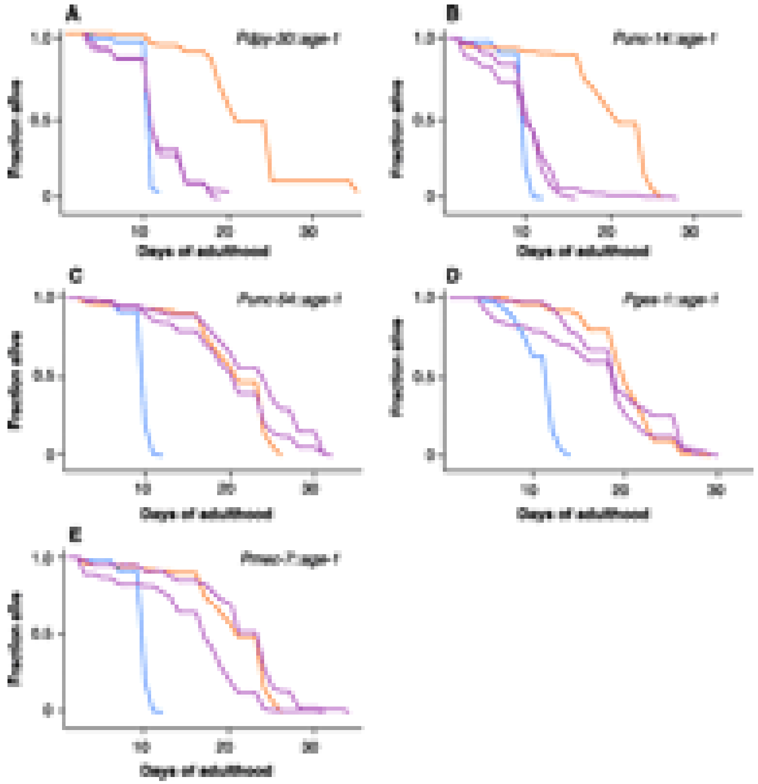
- Vertices: generators and substations
- Edges: high-voltage power transmission lines and transformers
- Line thickness and color indicate the voltage level
  - Red 765 kV, 500 kV; brown 345 kV; green 230 kV; grey 138 kV



Case 3: C. Elegans Nervous System

- Vertices: neurons in the C. elegans worm
- Edges: axons/synapses between neurons





Two More Examples

- M. Newman on scientific collaboration networks
  - coauthorship networks in several distinct communities
  - differences in degrees (papers per author)
  - empirical verification of
    - giant components
    - small diameter (mean distance)
    - high clustering coefficient
- Alberich et al. on the Marvel Universe
  - *purely fictional* social network
  - two characters linked if they appeared together in an issue
  - "empirical" verification of
    - heavy-tailed distribution of degrees (issues and characters)
    - giant component
    - rather *small* clustering coefficient

One More (Structural) Property...

- A properly tuned α-model can *simultaneously* explain
  - small diameter
  - high clustering coefficient
- But what about heavy-tailed degree distributions?
  - α-model and simple variants will *not* explain this
  - intuitively, no “bias” towards large degree evolves
  - all vertices are created equal
- Can concoct many *bad* generative models to explain
  - generate NW according to Erdos-Renyi, reject if tails not heavy
  - describe *fixed* NWs with heavy tails
    - all connected to v1; N/2 connected to v2; etc.
    - not clear we can get a precise power law
    - not modeling *variation*
  - why would the world evolve this way?
- As always, we want a “natural” model

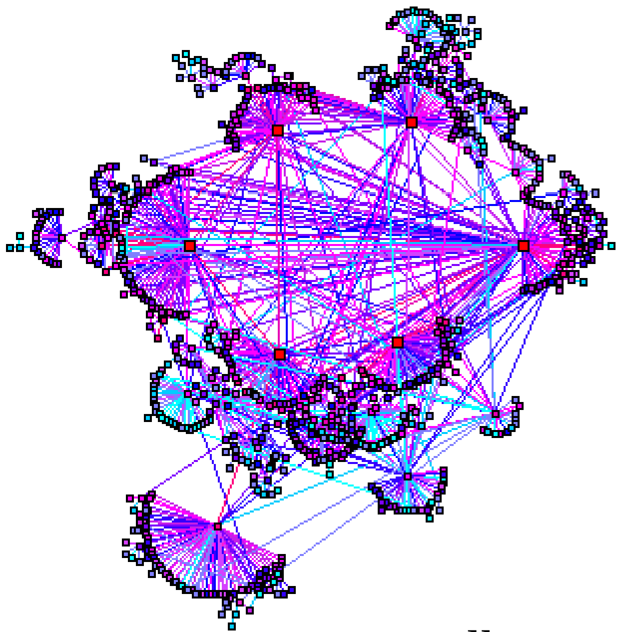
Models of Social Network Generation

- Random Graphs (Erdös-Rényi models)
- Watts-Strogatz models
- Scale-free Networks

World Wide Web

Nodes: WWW documents
Links: URL links

800 million documents
(S. Lawrence, 1999)

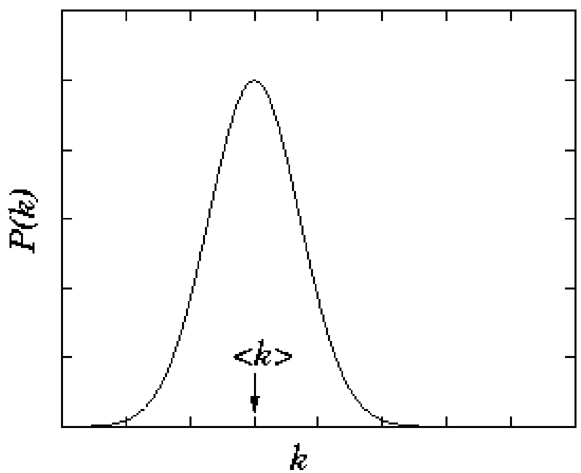


ROBOT: **collects all URL's found in a document and follows them recursively**

R. Albert, H. Jeong, A-L Barabasi, Nature, **401** 130 (1999)

World Wide Web

Expected Result



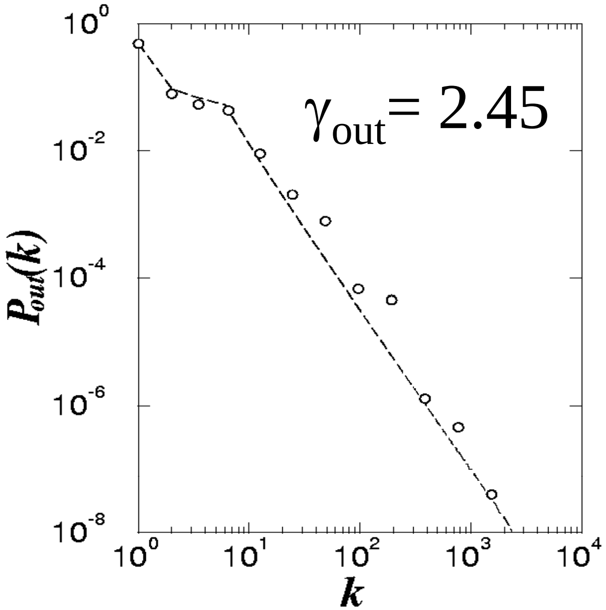
$\langle k \rangle \sim 6$

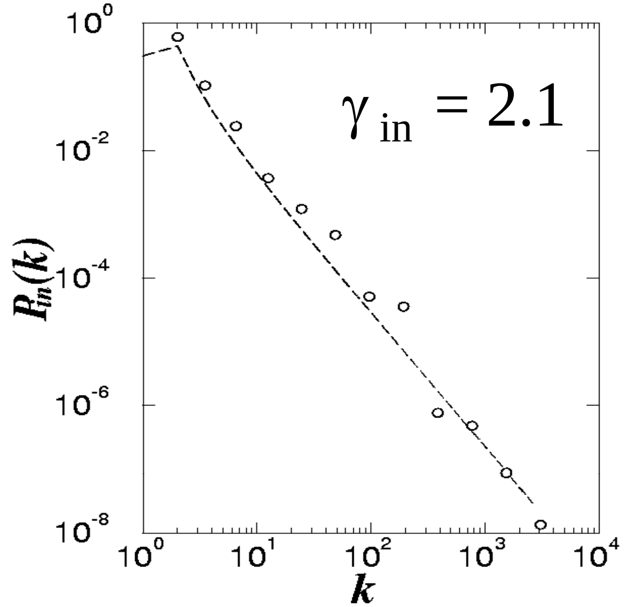
$P(k=500) \sim 10^{-99}$

$N_{WWW} \sim 10^{9}$

$\Rightarrow N(k=500) \sim 10^{-90}$

Real Result





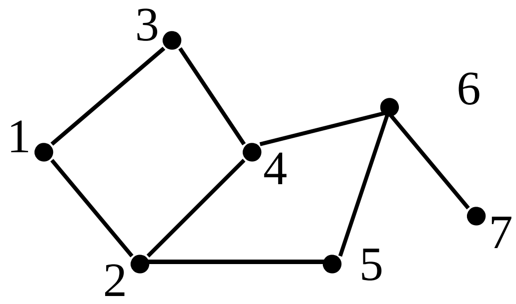
$P_{out}(k) \sim k^{-\gamma out}$ $P_{in}(k) \sim k^{-\gamma in}$

$P(k=500) \sim 10^{-6}$ $N_{WWW} \sim 10^{9}$

$\Rightarrow N(k=500) \sim 10^{3}$

J. Kleinberg, et. al, Proceedings of the ICCC (1999)

World Wide Web



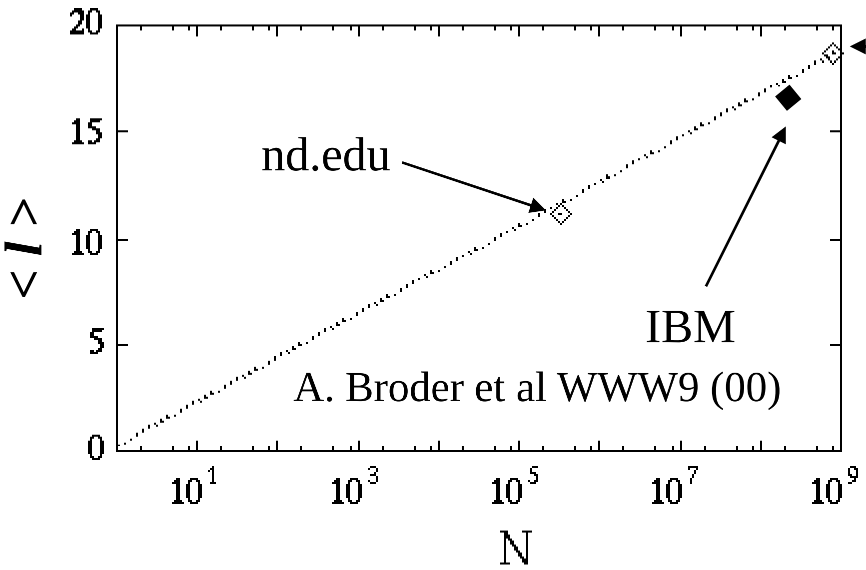
l_{15}=2 [1→2→5]

l_{17}=4 [1→3→4→6 → 7]

… **< *l* > = ??**

- **Finite size scaling**: create a network with N nodes with $P_{in}(k)$ and $P_{out}(k)$

< l > = 0.35 + 2.06 log(N)

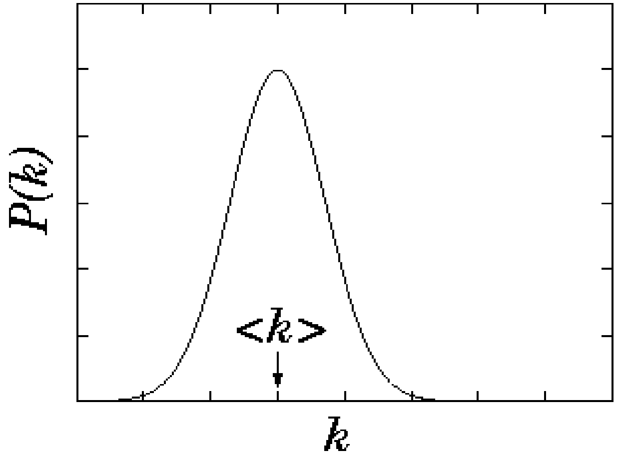


19 degrees of separation
R. Albert et al Nature (99)

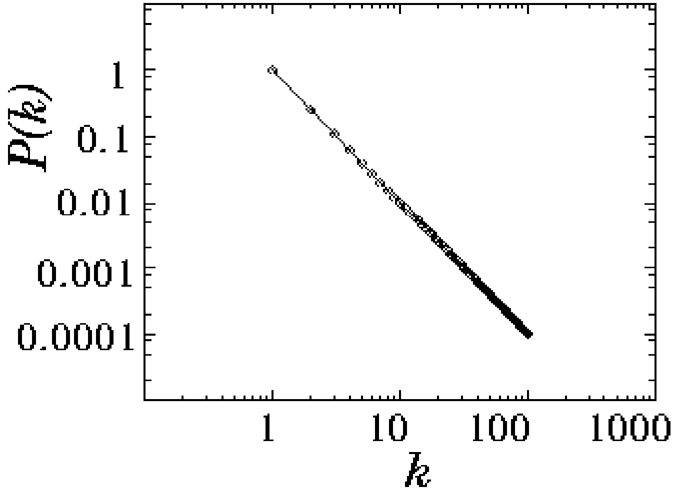
based on 800 million webpages
[S. Lawrence et al Nature (99)]

What does that mean?

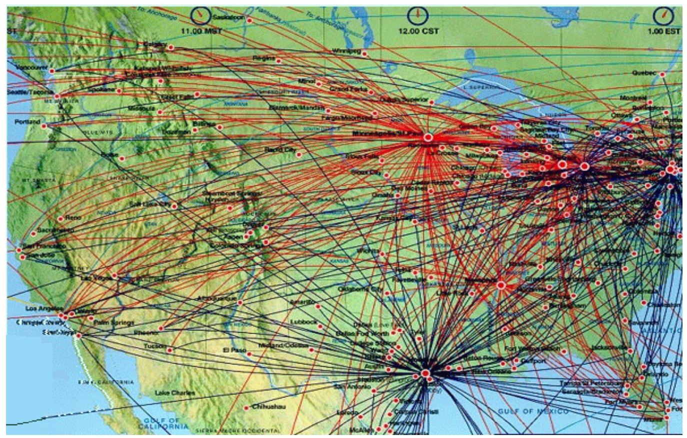
Poisson distribution



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

- Start with (say) two vertices connected by an edge
- For i = 3 to N:
  - for each 1 <= j < i, d(j) = degree of vertex j so far
  - let Z = S d(j) (sum of all degrees so far)
  - add new vertex i with k edges back to {1, ..., i-1}:
    - i is connected back to j with probability d(j)/Z
- Vertices j with high degree are likely to get more links!
- “Rich get richer”
- Natural model for many processes:
  - hyperlinks on the web
  - new business and social contacts
  - transportation networks
- Generates a power law distribution of degrees
  - exponent depends on value of k

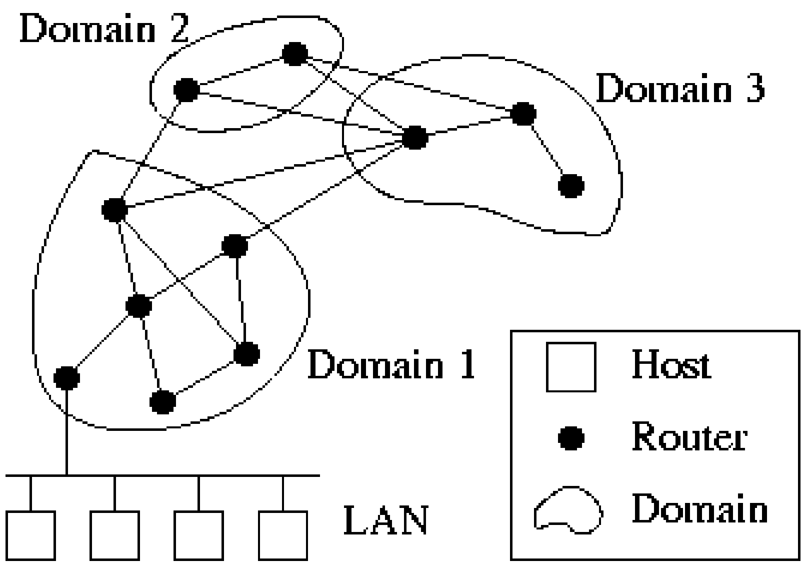
Scale-Free Networks

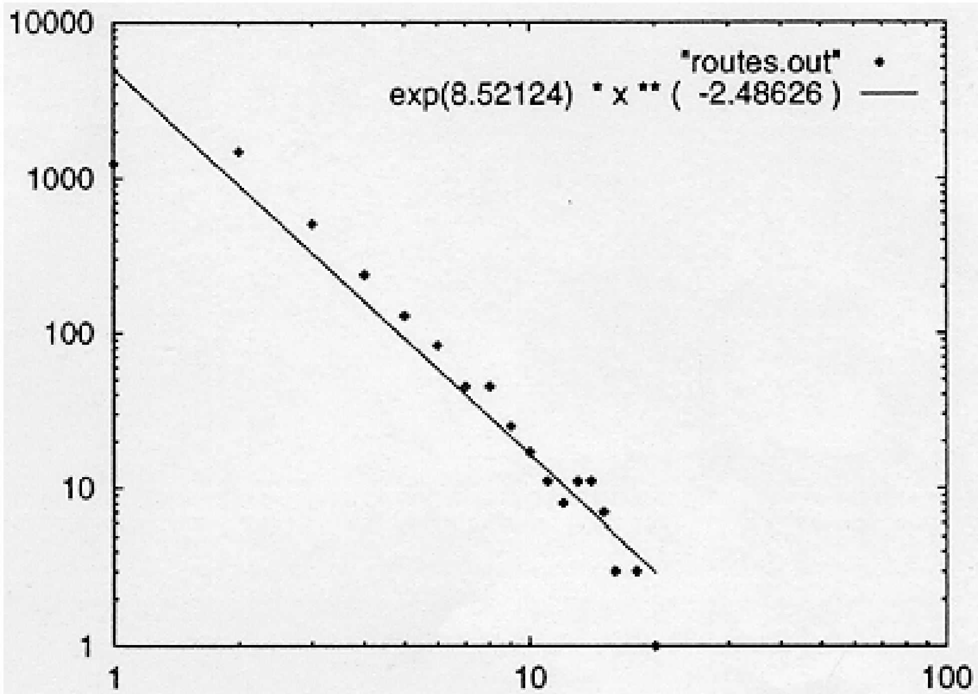
- Preferential attachment explains
  - heavy-tailed degree distributions
  - small diameter (~log(N), via “hubs”)

- Will *not* generate high clustering coefficient
  - no bias towards local connectivity, but towards hubs

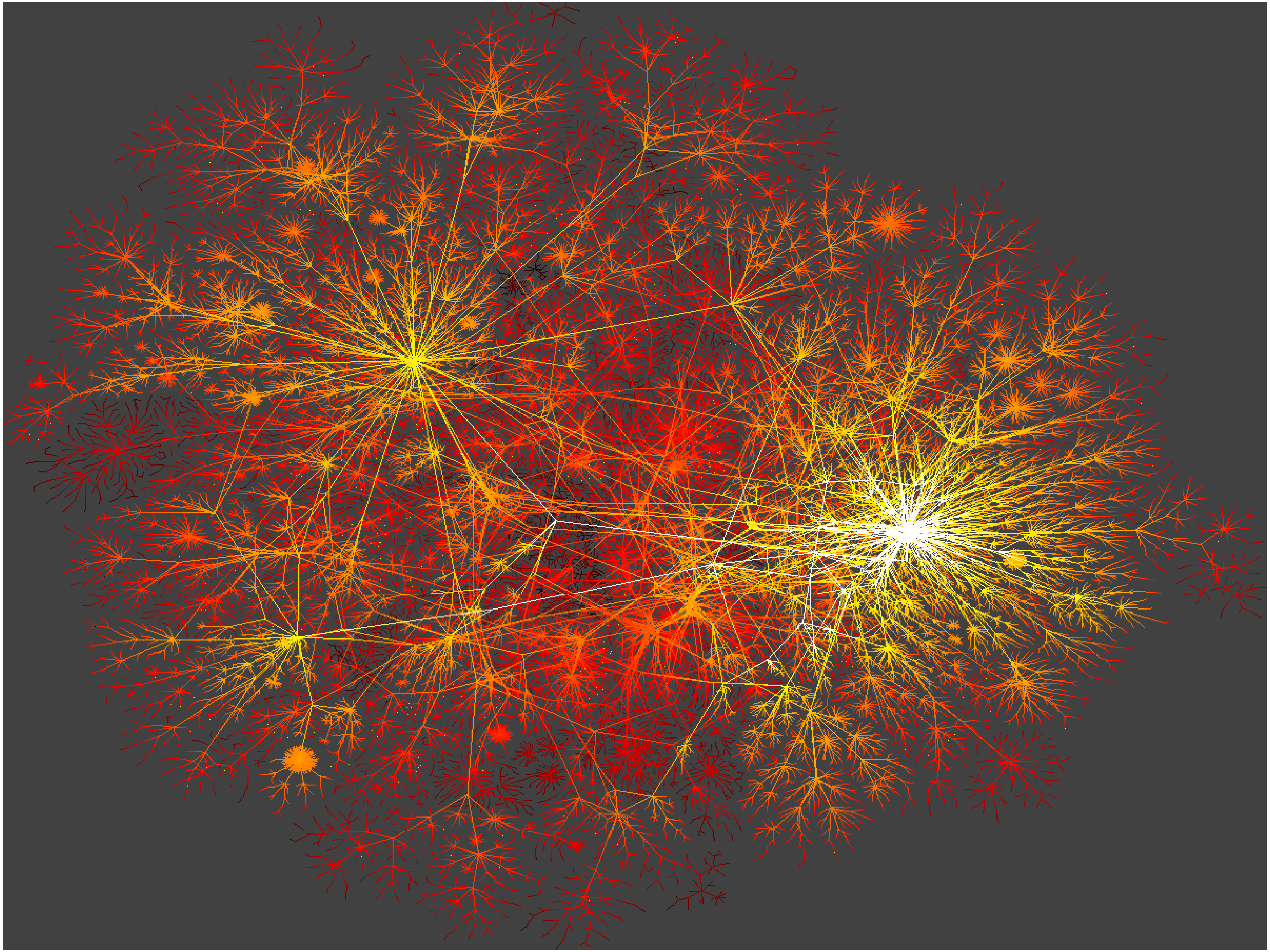
Case1: Internet Backbone

Nodes: computers, routers
Links: physical lines





(Faloutsos, Faloutsos and Faloutsos, 1999)



Case2: Actor Connectivity

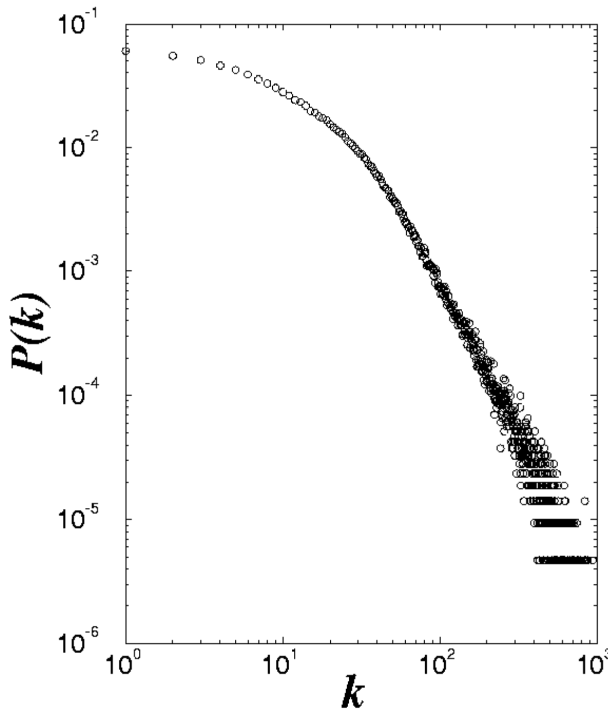
Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)

N = 212,250 actors
⟨k⟩ = 28.78

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

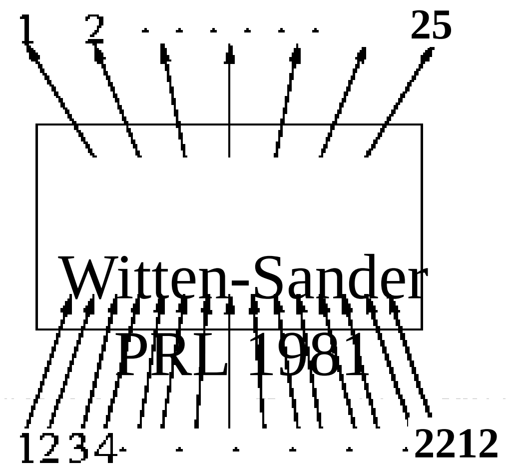
γ=2.3

Nodes: actors
Links: cast jointly

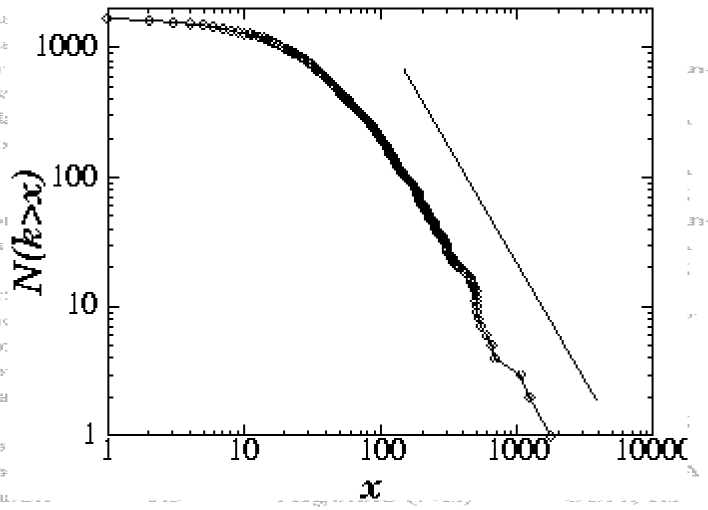


Case 3: Science Citation Index

Nodes: papers
Links: citations



1736 PRL papers (1988)



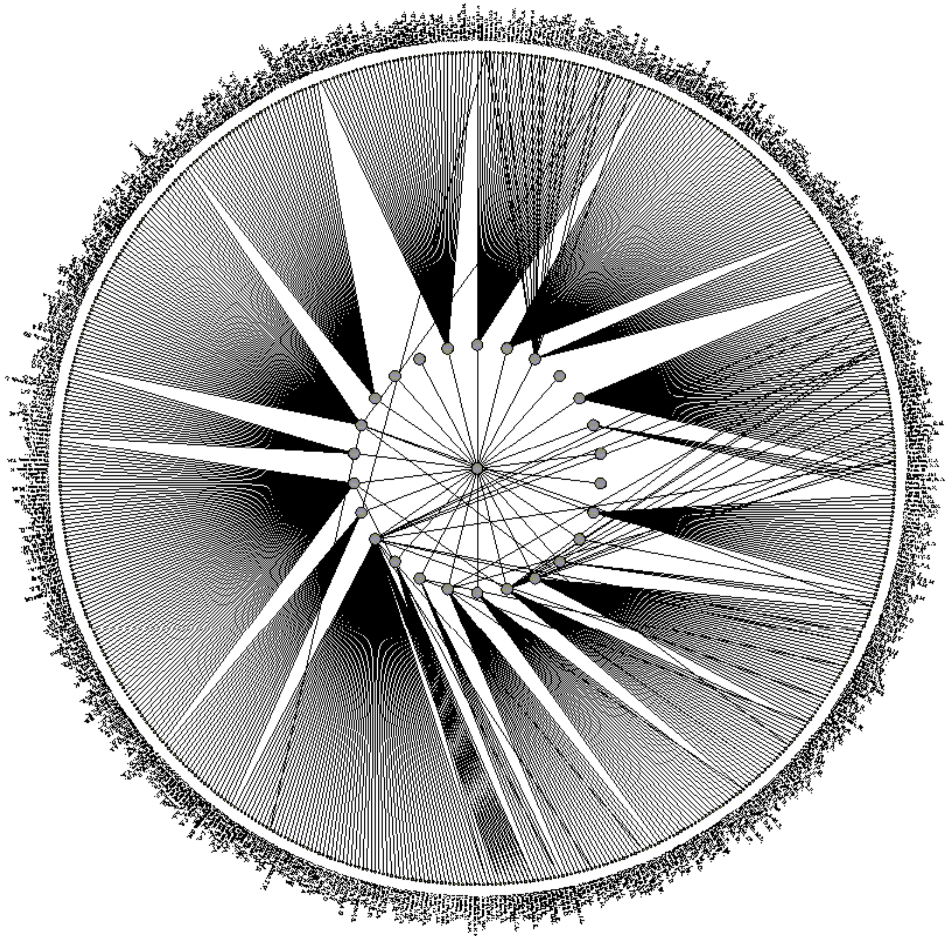
$P(k) \sim k^{-\gamma}$

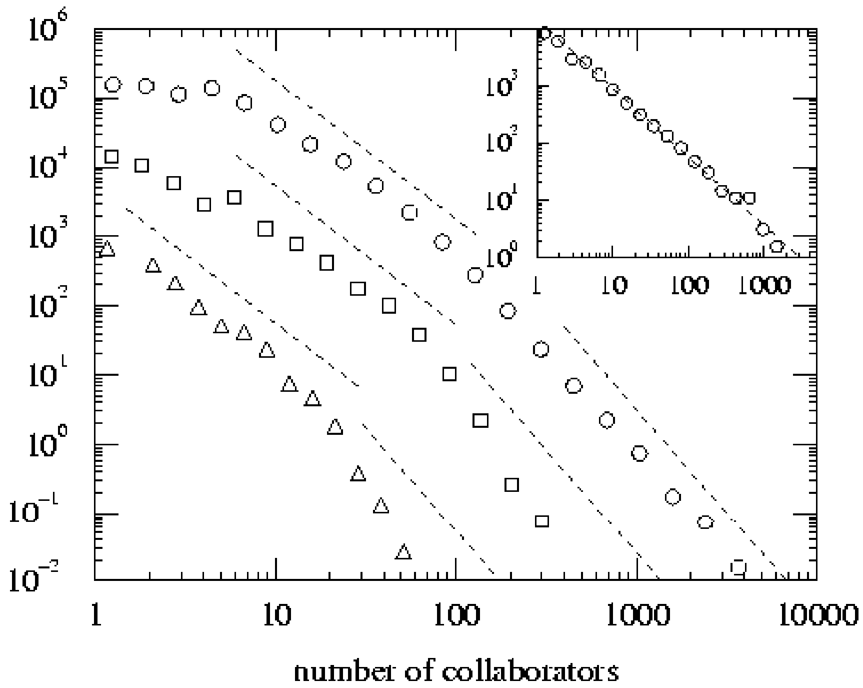
$(\gamma = 3)$

(S. Redner, 1998)

Case 4: Science Coauthorship

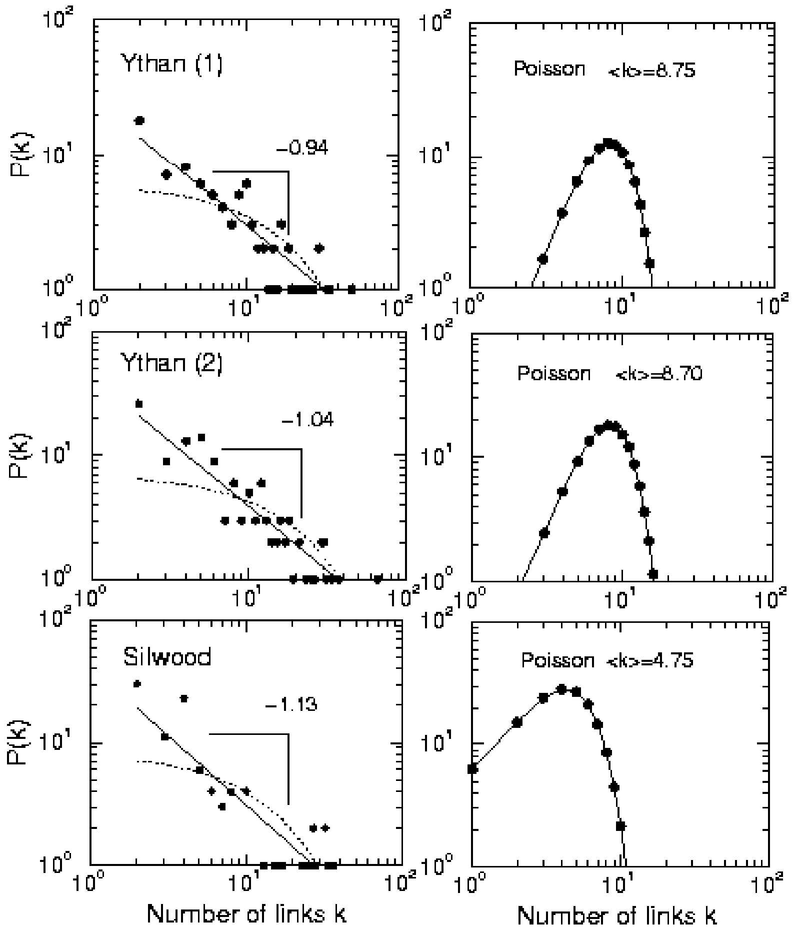
Nodes: scientist (authors)
Links: write paper together





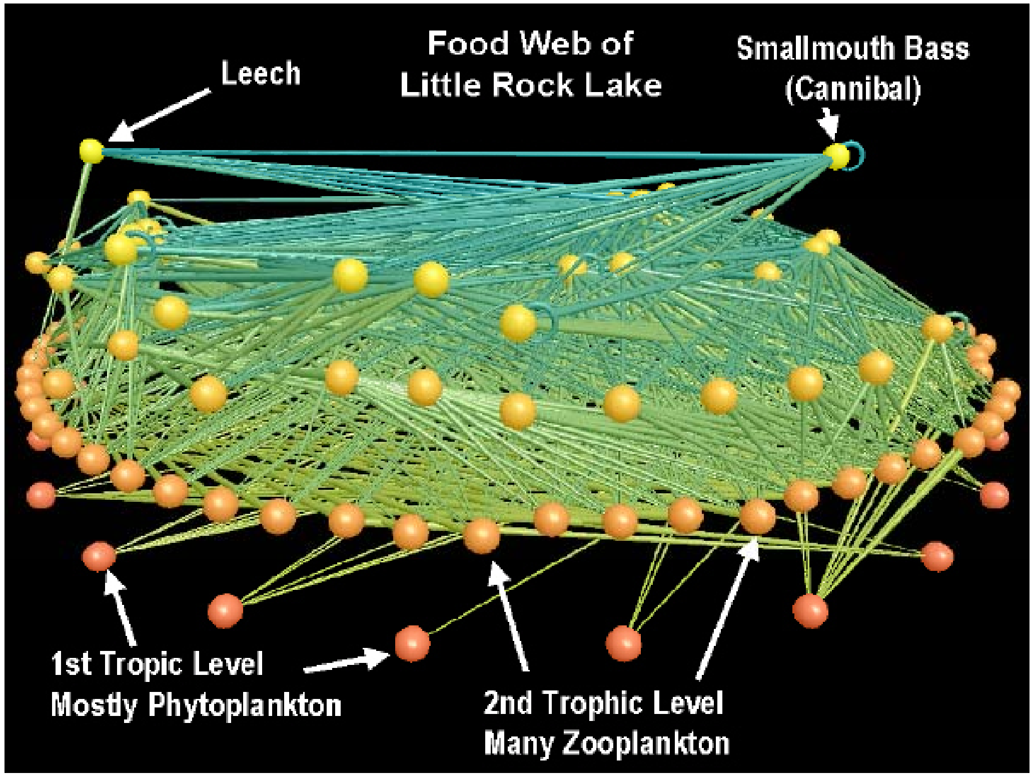
(Newman, 2000, H. Jeong et al 2001)

Case 5: Food Web



R. Sole (cond-mat/0011195)

Nodes: trophic species
Links: trophic interactions

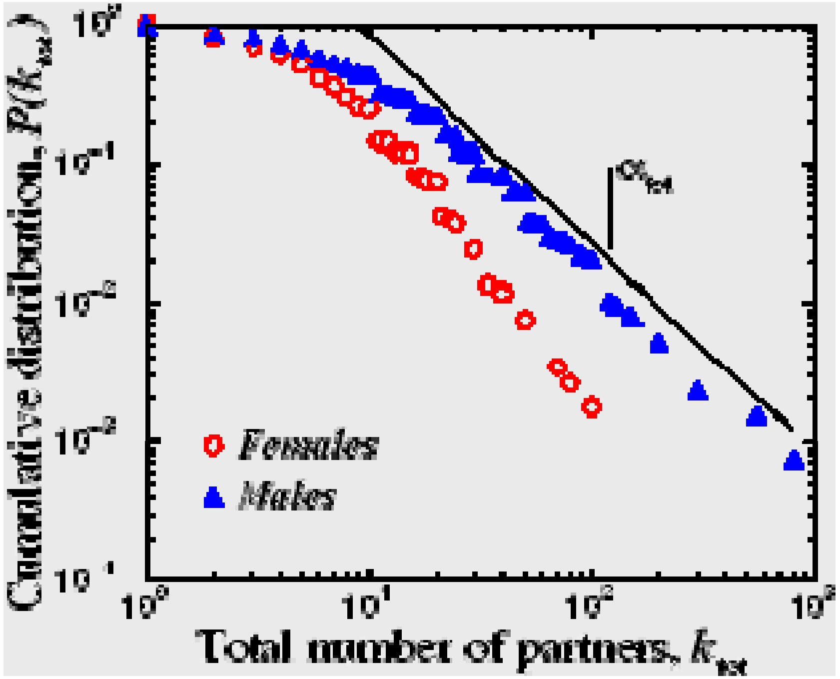


R.J. Williams, N.D. Martinez *Nature* (2000)

Case 6: Sex-Web

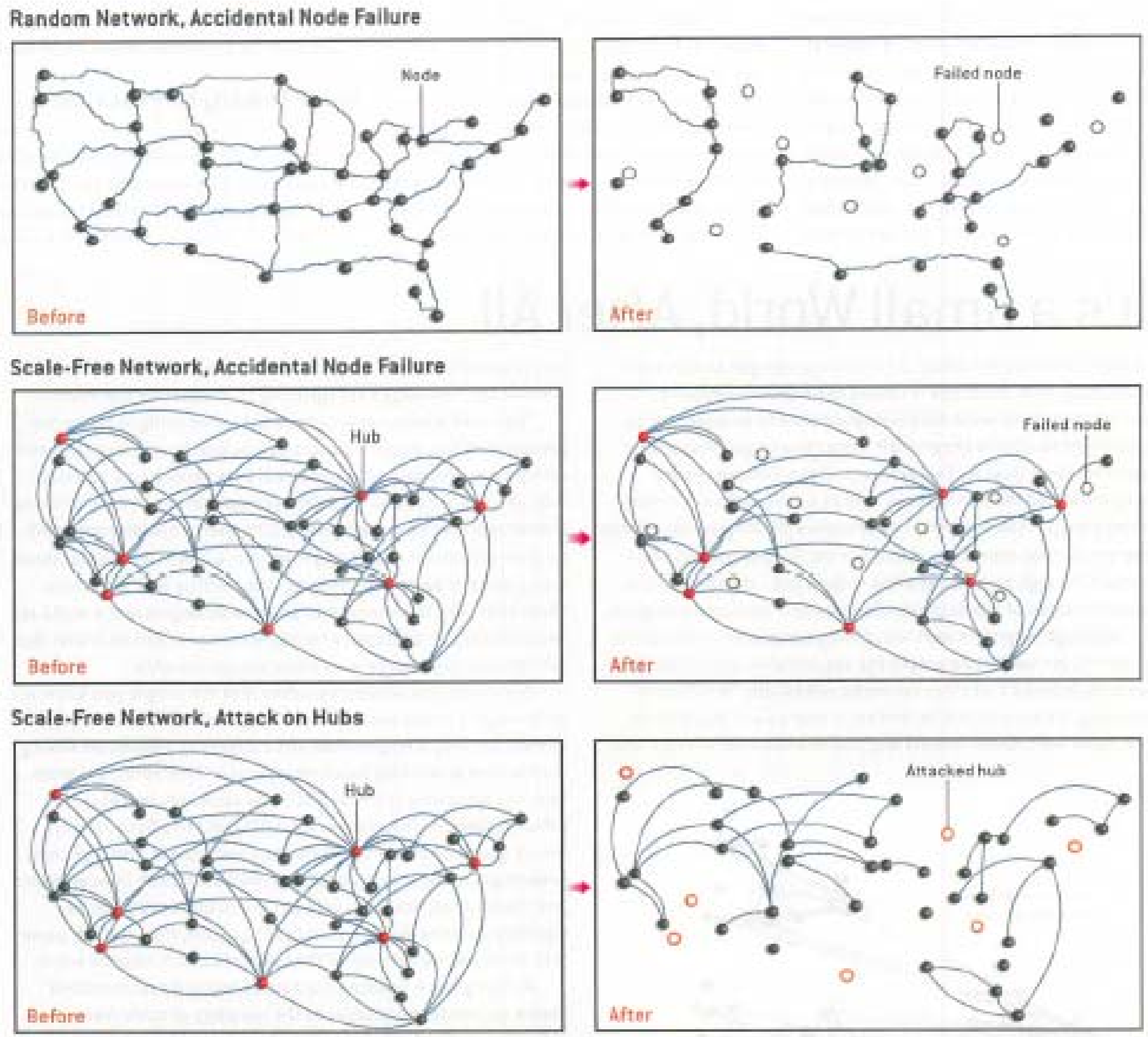
Nodes: people (Females; Males)

Links: sexual relationships



4781 Swedes; 18-74;
59% response rate.
Liljeros et al. Nature 2001

Robustness of Random vs. Scale-Free Networks



- 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.