Data Mining: Concepts and Techniques

— Chapter 9 —

9.2. Social Network Analysis

Jiawei Han and Micheline Kamber

Department of Computer Science

University of Illinois at Urbana-Champaign

www.cs.uiuc.edu/~hanj

Acknowledgements: Based on the slides by Sangkyum Kim and Chen Chen

Social Network Analysis

- Social Networks: An Introduction
- Primitives for Network Analysis
- Different Network Distributions
- Models of Social Network Generation
- Mining on Social Network
- Summary

Social Networks

- Social network: A social structure made of nodes (individuals or organizations) that are related to each other by various interdependencies like friendship, kinship, like, ...
- Graphical representation
  - Nodes = members
  - Edges = relationships
- Examples of typical social networks on the Web
  - Social bookmarking (Del.icio.us)
  - Friendship networks (Facebook, Myspace, LinkedIn)
  - Blogosphere
  - Media Sharing (Flickr, Youtube)
  - Folksonomies

Society

Nodes: individuals

Links: social relationship (family/work/friendship/etc.)

S. Milgram (1967)

John Guare

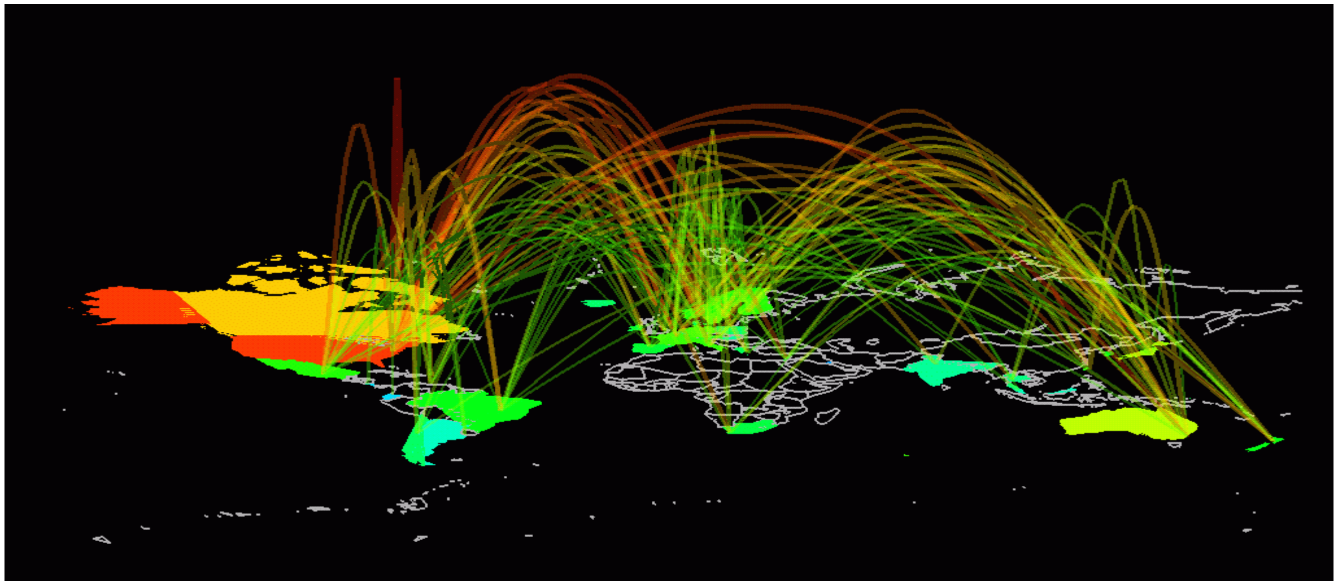
Six Degrees of Separation

Social networks: Many *individuals* with *diverse social interactions* between them.

Communication networks

The earth is developing an electronic nervous system, a network with diverse **nodes** and **links** are

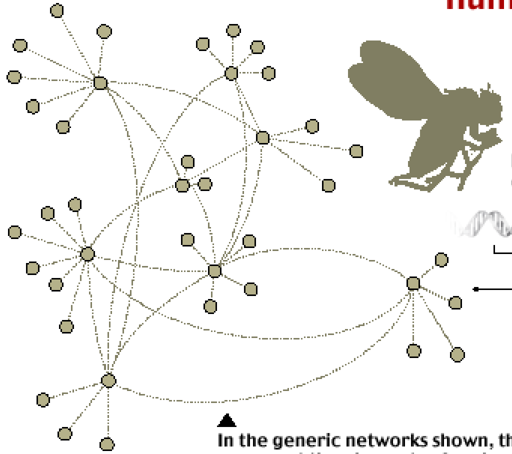
nodes	links
-computers	-phone lines
-routers	-TV cables
-satellites	-communication lines



Communication networks: Many non-identical components with diverse connections between them

Humans have only about three times as many genes as the fly,

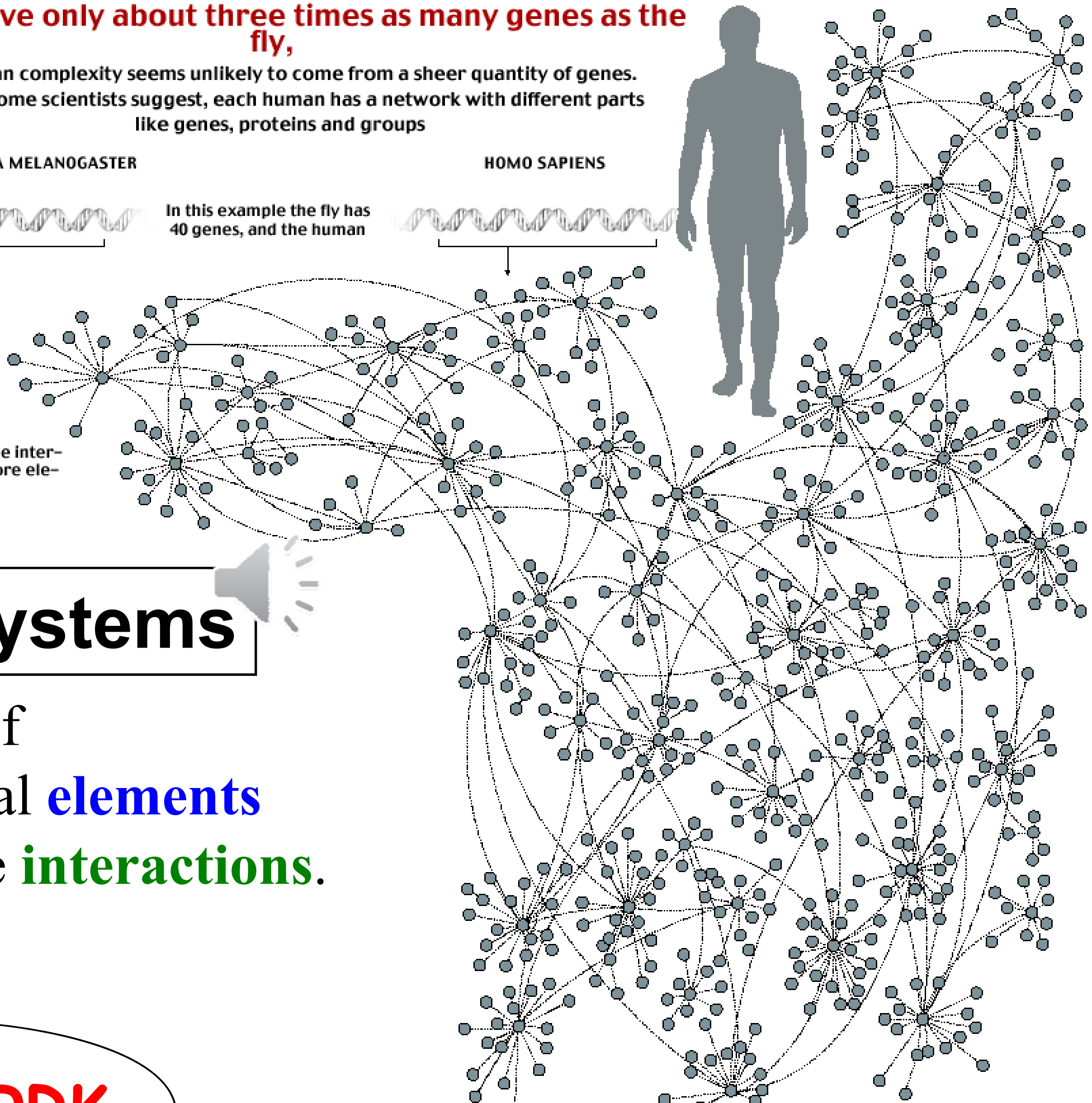
so human complexity seems unlikely to come from a sheer quantity of genes. Rather, some scientists suggest, each human has a network with different parts like genes, proteins and groups



DROSOPHILA MELANOGASTER
(Fruit fly)

In this example the fly has 40 genes, and the human

HOMO SAPIENS



In the generic networks shown, the points represent the elements of each organism's genetic network, and the dotted lines show the interactions between them. Humans have many more ele-

Sources: Dr. Albert-László Barabási, University of Notre Dame; Science; Celera Genomics

Complex systems

Made of
many non-identical **elements**
connected by diverse **interactions**.

⇩

NETWORK

“Natural” Networks and Universality

- Consider many kinds of networks:
  - social, technological, business, economic, content,…
- These networks tend to share certain *informal* properties:
  - Large scale; continual growth
  - Distributed and organic growth: vertices “decide” who to link to
  - Mixture of local and long-distance connections
    - Abstract notions of distance: geographical, content, social,…
- Questions:
    - Do natural networks share more *quantitative* universals?
    - What would these “universals” be?
    - How can we measure them?
- This is the domain of *social network theory* or *link analysis*

Social Network Analysis

- Social Networks: An Introduction
- Primitives for Network Analysis
- Different Network Distributions
- Models of Social Network Generation
- Mining on Social Network
- Summary

Networks and Their Representations

- A network (or a graph): G = (V, E), where V: vertices (or nodes), and E: edges (or links)
  - Multi-edge: if more than one edge between the same pair of vertices
  - Self-edge (self-loop): if an edge connects vertex to itself
- Simple network/graph if a network has neither self-edges nor multi-edges
- Adjacency matrix:
  - A_{ij} = 1 if there is an edge between vertices i and j; 0 otherwise
- Weighted networks:
  - Edges having weight (strength), usually a real number
- Directed network (directed graph): if each edge has a direction
  - A_{ij} = 1 if there is an edge from i to j; 0 otherwise

Cocitation and Bibliographic Coupling

- Cocitation of vertices i and j: # of vertices having outgoing edges pointing to both i and j

$$A_{ik}A_{jk} = 1 \text{ if } i \text{ and } j \text{ are both cited by } k$$

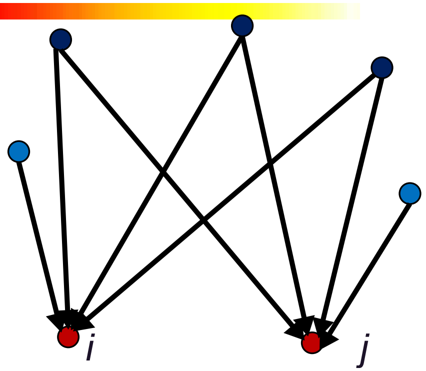
  - Cocitation of i and j:

$$C_{ij} = \Sigma_{k=1}^{n} A_{ik}A_{jk} = \Sigma_{k=1}^{n} A_{ik}A_{kj}^{T}$$

  - Cociation matrix: It is a symmetric matrix

$$\mathbf{C} = \mathbf{A}\mathbf{A}^{T}$$

  - Diagonal matrix (C_{ii}): total # papers citing i



Vertices i and j are co-cited by 3 papers

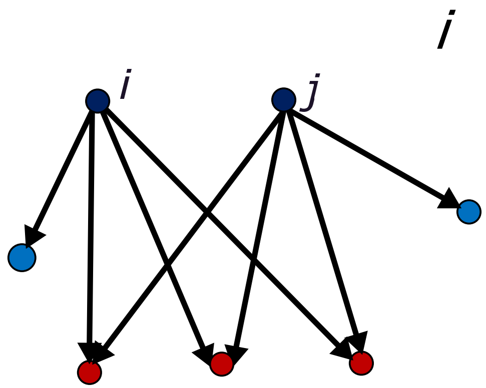
- Bibliographic coupling of vertices i and j: # of other vertices to which both point

$$A_{ki}A_{kj} = 1 \text{ if } i \text{ and } j \text{ both cite } k$$

  - Bibliographic coupling of i and j:

$$B_{ij} = \Sigma_{k=1}^{n} A_{ki}A_{kj} = \Sigma_{k=1}^{n} A_{ik}^{T}A_{kj}$$

  - Cociation matrix: $\mathbf{B} = \mathbf{A}^{T}\mathbf{A}$
  - Diagonal matrix (B_{ii}): total # papers cited by i



Vertices i and j cite 3 same papers

Cocitation & Bibliographic Coupling: Comparison

- Two measures are affected by the number of incoming and outgoing edges that vertices have
- For strong cocitation: must have a lot of incoming edges
  - Must be well-cited (influential) papers, surveys, or books
  - Takes time to accumulate citations
- Strong bib-coupling if two papers have similar citations
  - A more uniform indicator of similarity between papers
  - Can be computed as soon as a paper is published
  - Not change over time
- Recent analysis algorithms
  - HITS explores both cocitation and bibliographic coupling

Degree and Network Density

- Degree of a vertex *i*: $k_i = \sum_{j=1}^{n} A_{ij}$
- # of edges m = 1/2 of sum of degrees of all the vertices:

$$m = \frac{1}{2}\sum_{i}^{n} k_i = \frac{1}{2}\sum_{ij} A_{ij}$$

- The mean degree c of a vertex in an undirected graph:

$$c = \frac{1}{n}\sum_{i}^{n} k_i = \frac{2m}{n}$$

- Density ρ of a graph: $\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)} = \frac{c}{n-1}$
- A network is **dense** if density ρ tends to be a constant as n → ∞
- A network is **sparse** if density ρ → 0 as n → ∞. The fraction of nonzero element in the adjacency matrix tends to zero
- Internet, WWW and friendship networks are usually regarded as sparse

Social Network Analysis

- Social Networks: An Introduction
- Primitives for Network Analysis
- Different Network Distributions
- Models of Social Network Generation
- Mining on Social Network
- Summary

Some Interesting Quantities

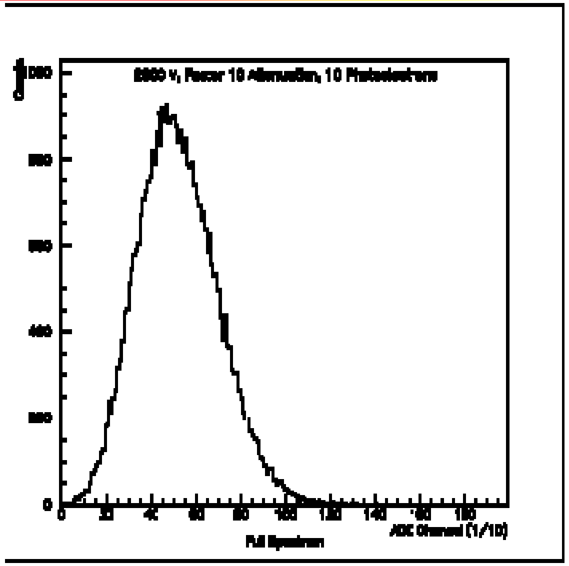
- *Connected components:*
  - how many, and how large?
- *Network diameter:*
  - maximum (worst-case) or average?
  - exclude infinite distances? (disconnected components)
  - the small-world phenomenon
- *Clustering:*
  - to what extent links tend to cluster "locally"?
  - what is the balance between local and long-distance connections?
  - what roles do the two types of links play?
- *Degree distribution:*
  - what is the typical degree in the network?
  - what is the overall distribution?

A “Canonical” Natural Network has…

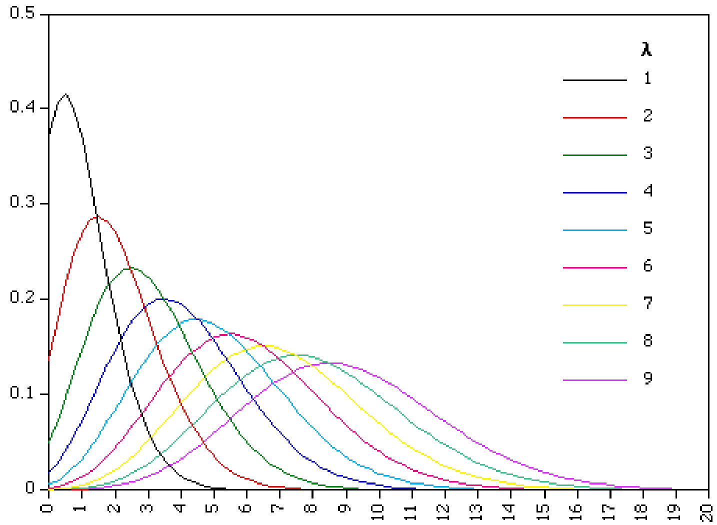
- *A few* connected components:
  - often only 1 or a small number, indep. of network size
- *Small* diameter:
  - often a constant independent of network size (like 6)
  - or perhaps growing only logarithmically with network size or even shrink?
    - typically exclude infinite distances
- A *high* degree of clustering:
  - considerably more so than for a random network
  - Related to small diameter
- A *heavy-tailed* degree distribution:
  - a small but reliable number of *high-degree vertices*
  - often of *power law* form

The Poisson Distribution

- Applies to variables taken on integer values > 0
- Often used to model *counts* of events
  - number of phone calls placed in a given time period
  - number of times a neuron fires in a given time period
- Single free parameter λ, probability of exactly *x* events:
  - $\exp(-\lambda)\, \lambda^x / x!$
  - mean and variance are both λ
- Binomial distribution with *n* large, $p = \lambda/n$ (λ fixed)
  - converges to Poisson with mean λ

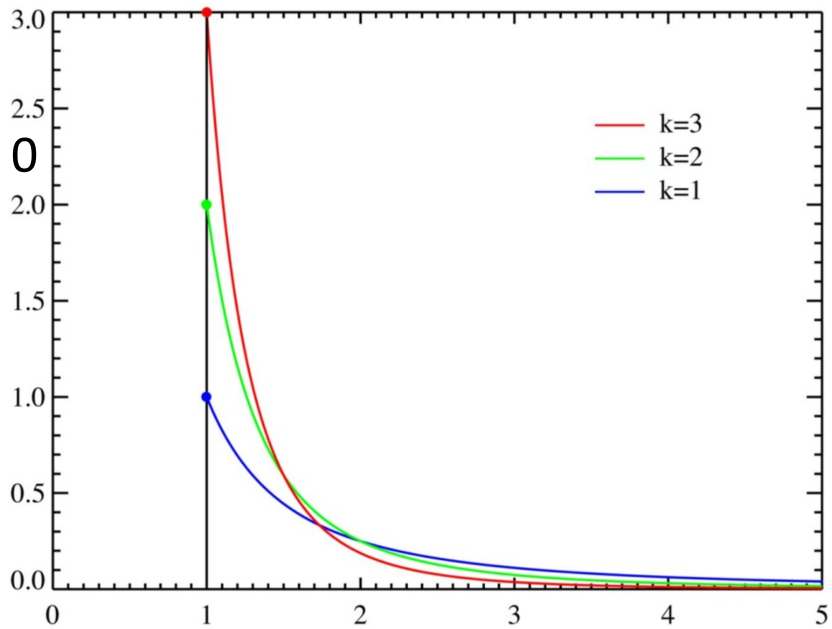


single photoelectron distribution



Power Law (or Pareto) Distributions

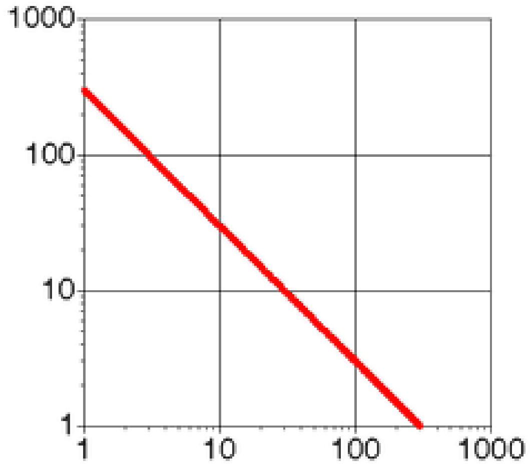
- Heavy-tailed, pareto, or *power law* distributions:
  - For variables assuming integer values > 0
  - probability of value x ~ $1/x^a$
  - Typically $0 < a < 2$; smaller a gives heavier tail
  - sometimes also referred to as being *scale-free*
- For Poisson distributions the tail probabilities approach 0 *exponentially* fast



- What kind of phenomena does this distribution model?
- What kind of process would *generate* it?

Distinguishing Distributions in Data

- All these distributions are *idealized models*
- In practice, we do not see distributions, but *data*
- Typical procedure to distinguish between Poisson, power law, …
  - might restrict our attention to a *range* of values of interest
  - accumulate *counts* of observed data into equal-sized bins
  - look at counts on a *log-log plot*
- power law:
  - $\log(\Pr[X = x]) = \log(1/x^a) = -a \log(x)$
  - linear, slope –a
- Poisson:
  - $\log(\Pr[X = x]) = \log(\exp(-l)\ l^x/x!)$
  - non-linear



Logarithmic scales on both axes

Social Network Analysis

- Social Networks: An Introduction
- Primitives for Network Analysis
- Different Network Distributions
- Models of Social Network Generation
- Mining on Social Network
- Summary

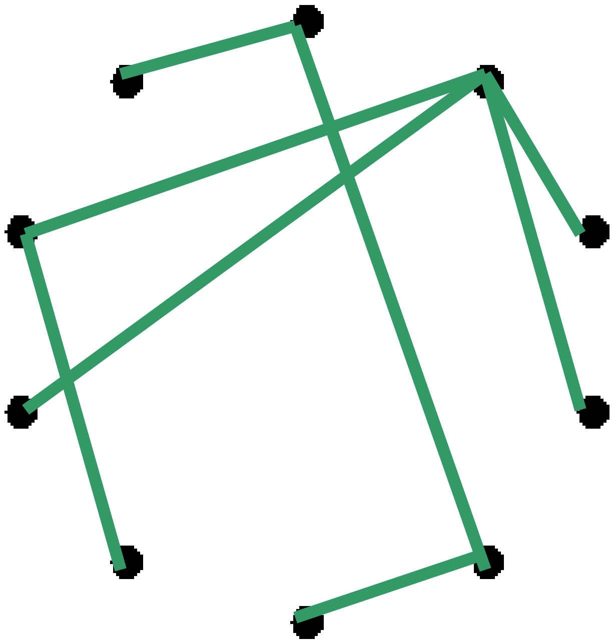
Models of Social Network Generation

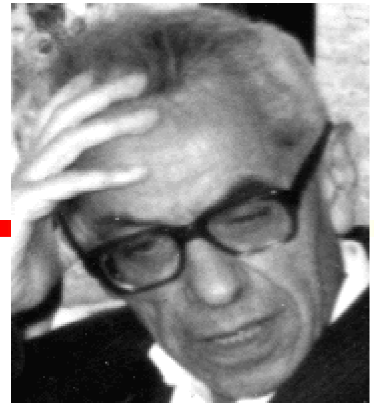
- Random Graphs (Erdös-Rényi models)
- Scale-free Networks

The Erdös-Rényi (ER) Model: A Random Graph Model

- A random graph is obtained by starting with a set of *N* vertices and adding edges between them at random
- Different *random graph models* produce different *probability distributions* on graphs
- Most commonly studied is the *Erdős–Rényi model*, denoted *G(N,p)*
  - **every possible edge occurs independently with probability *p***
- The usual *regime of interest* is when *p ~ 1/N, N is large*
  - 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*
  - *not* heavy-tailed

Erdös-Rényi Model (1959)

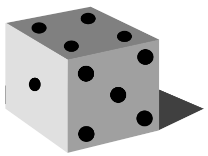




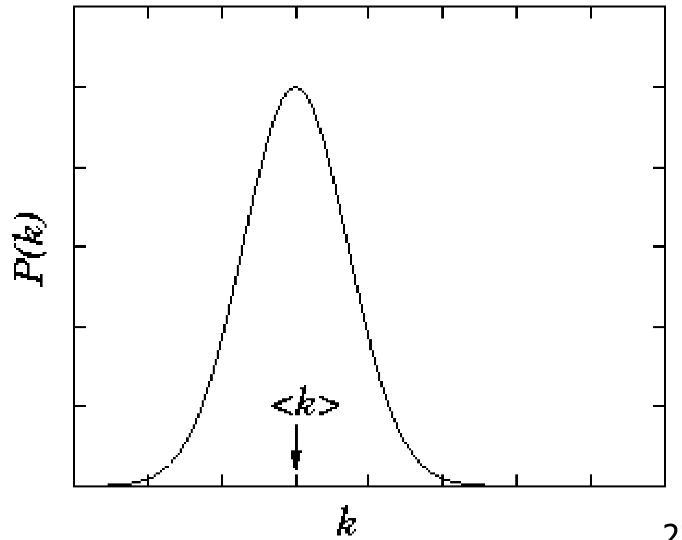
Pál Erdös
(1913-1996)

Connect with probability p

p=1/6
N=10
⟨k⟩~1.5



Poisson distribution



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

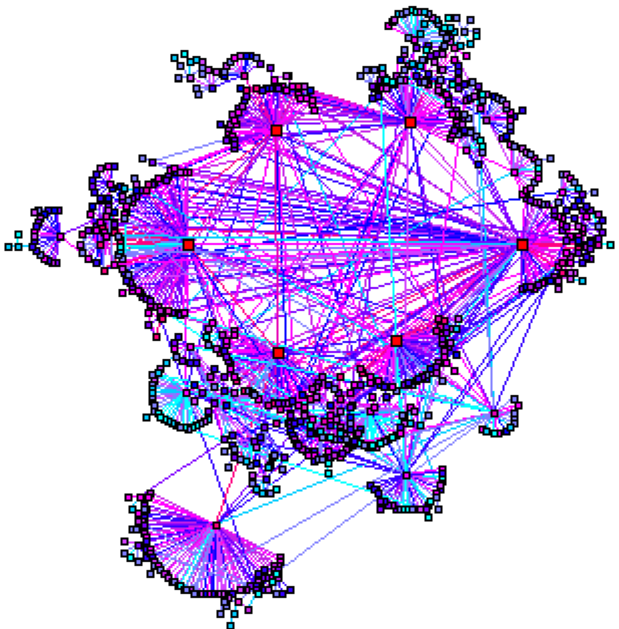
Models of Social Network Generation

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

World Wide Web

Nodes: WWW documents
Links: URL links

800 million documents
(S. Lawrence, 1999)

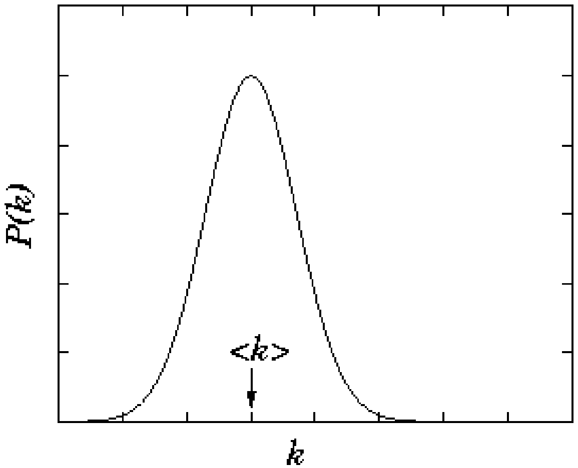


ROBOT: **collects all URL's found in a document and follows them recursively**

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

World Wide Web

Expected Result



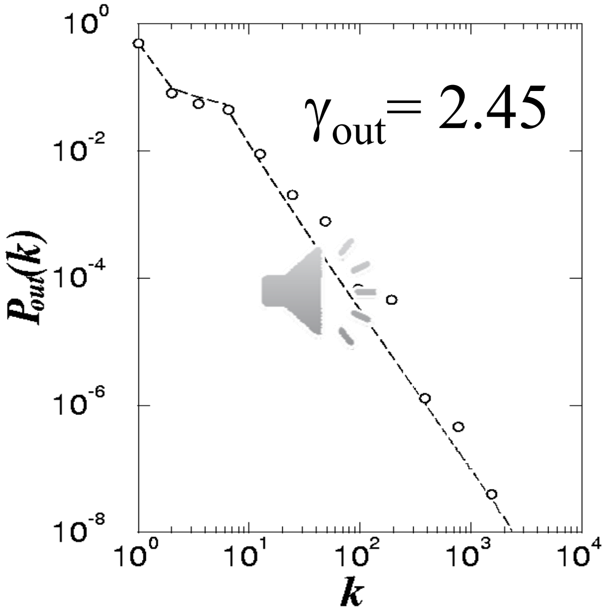
$\langle k \rangle \sim 6$

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

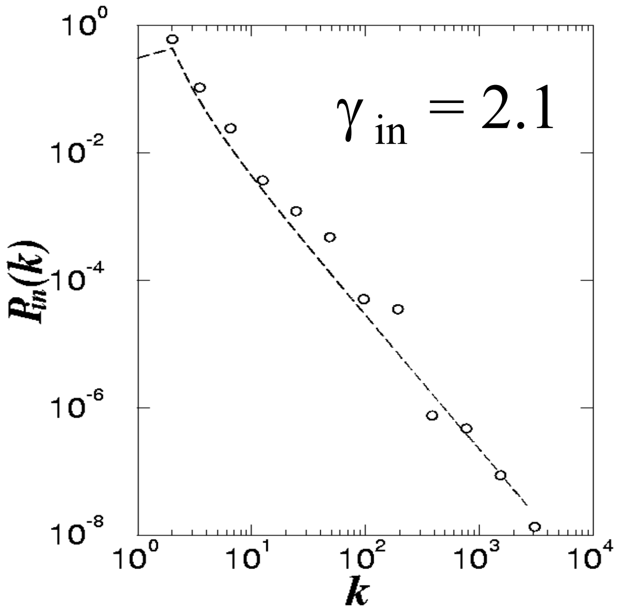
$N_{WWW} \sim 10^{9}$

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

Real Result



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



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

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

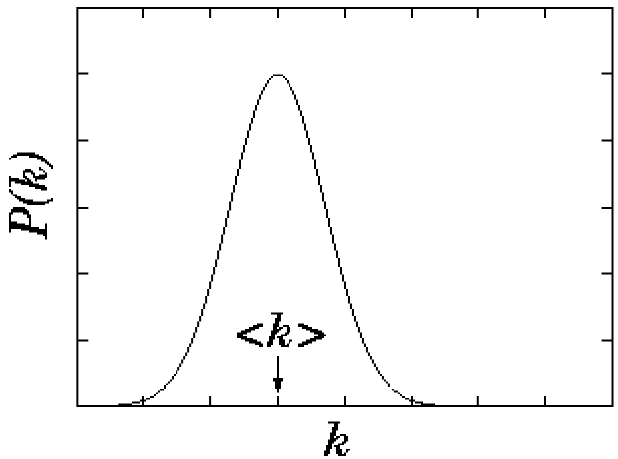
$N_{WWW} \sim 10^{9}$

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

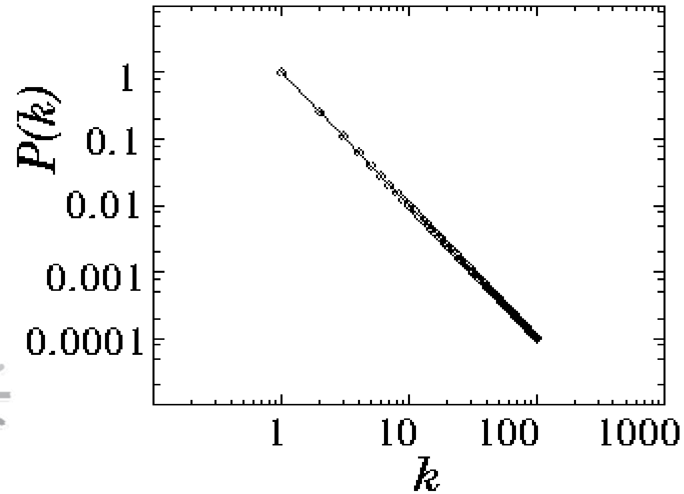
J. Kleinberg, et. al, Proceedings of the ICCC (1999)

What does that mean?

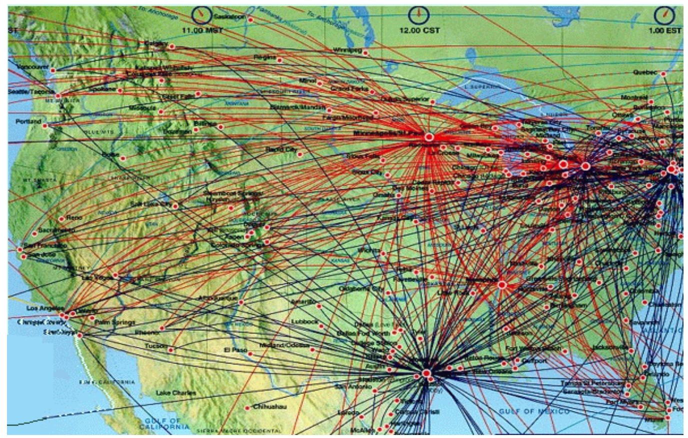
Poisson distribution



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 (random)
  - 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
- Generates a power law distribution of degrees
  - exponent depends on value of k

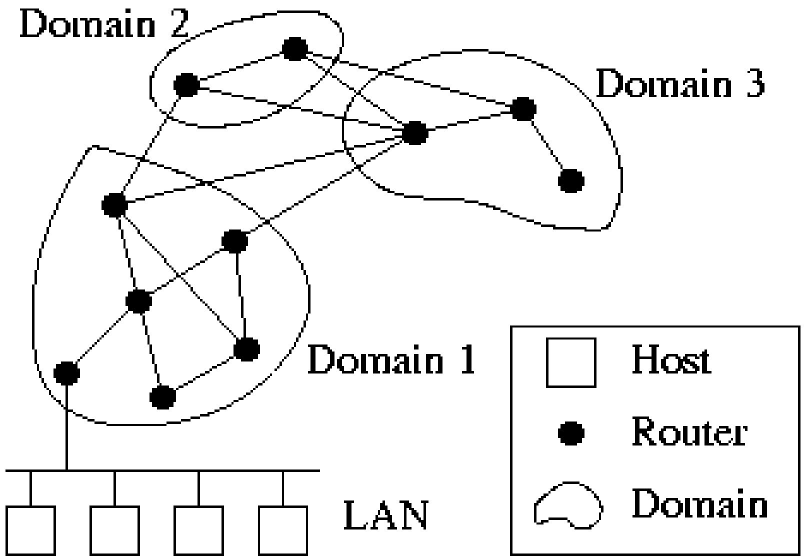
Scale-Free Networks

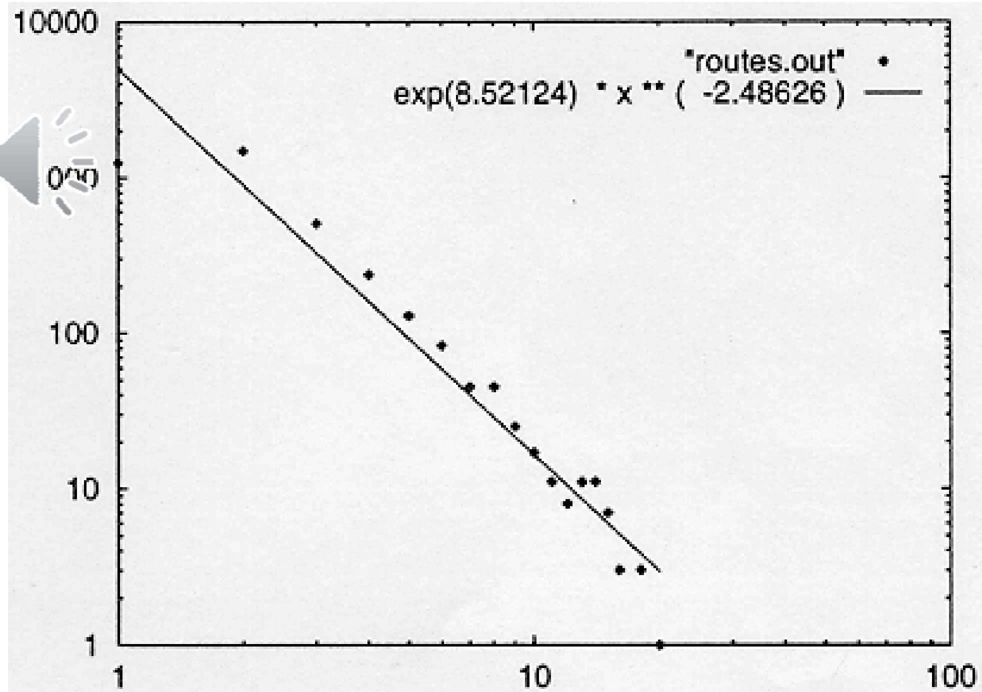
- Preferential attachment explains
  - heavy-tailed degree distributions
  - small diameter (~log(N), via "hubs")

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

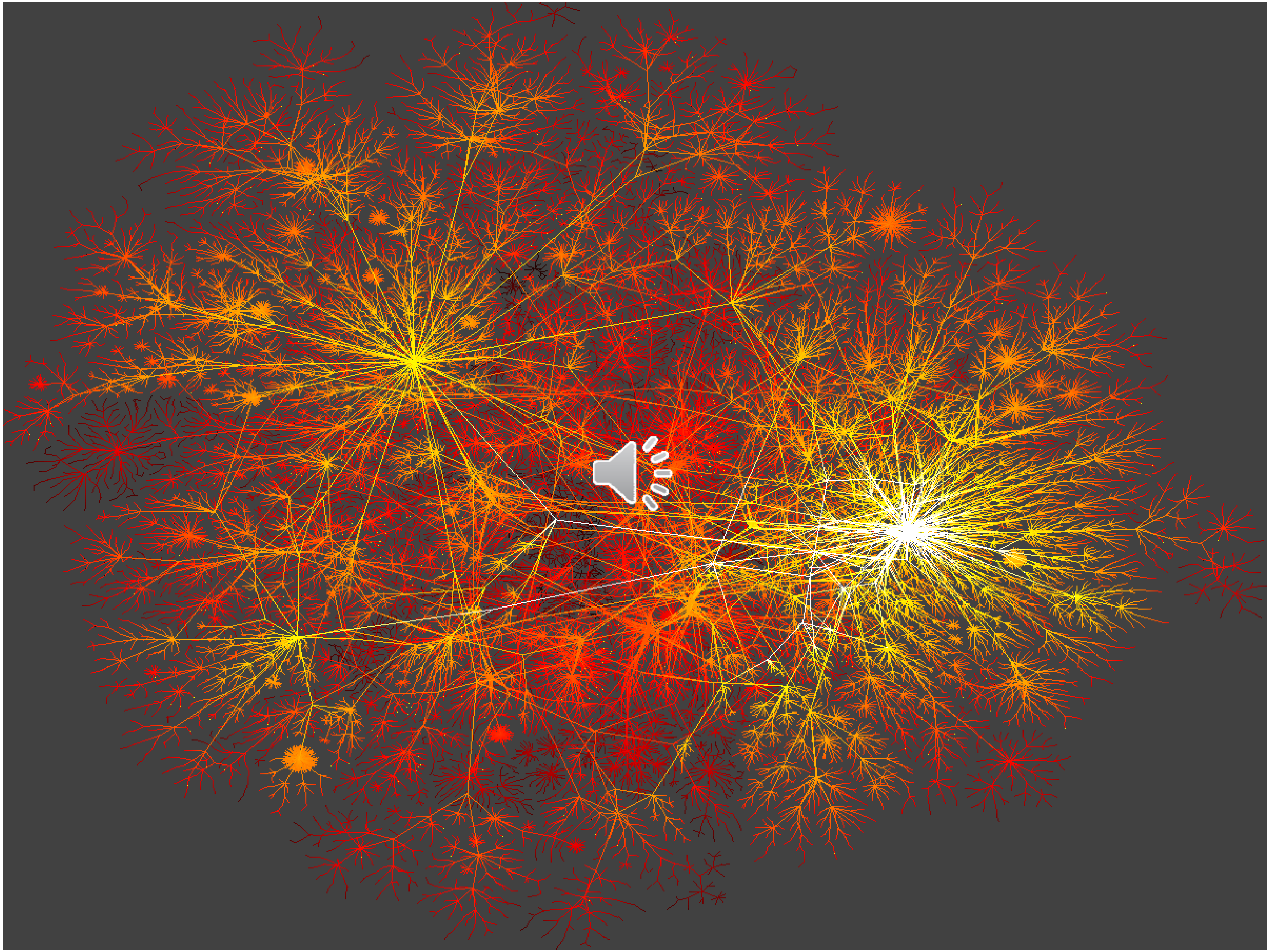
Case 1: Internet Backbone

Nodes: computers, routers
Links: physical lines





(Faloutsos, Faloutsos and Faloutsos, 1999)



Case 2: Actor Connectivity

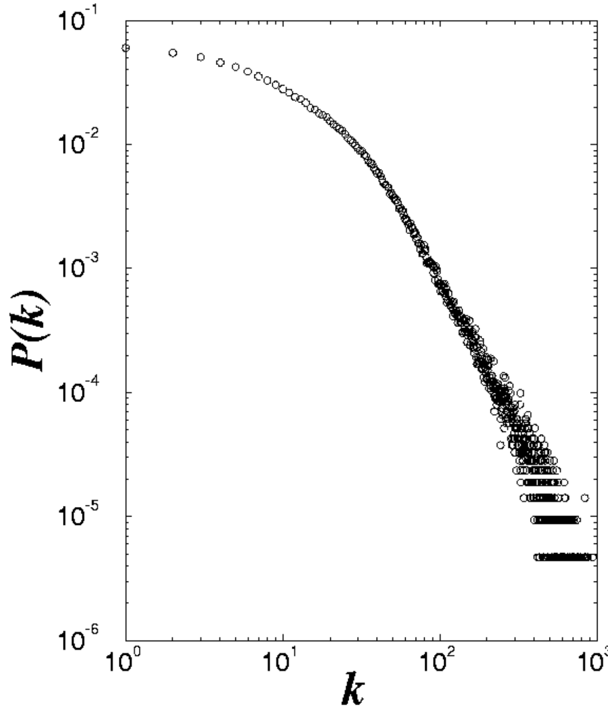
Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)

N = 212,250 actors
$\langle k \rangle = 28.78$

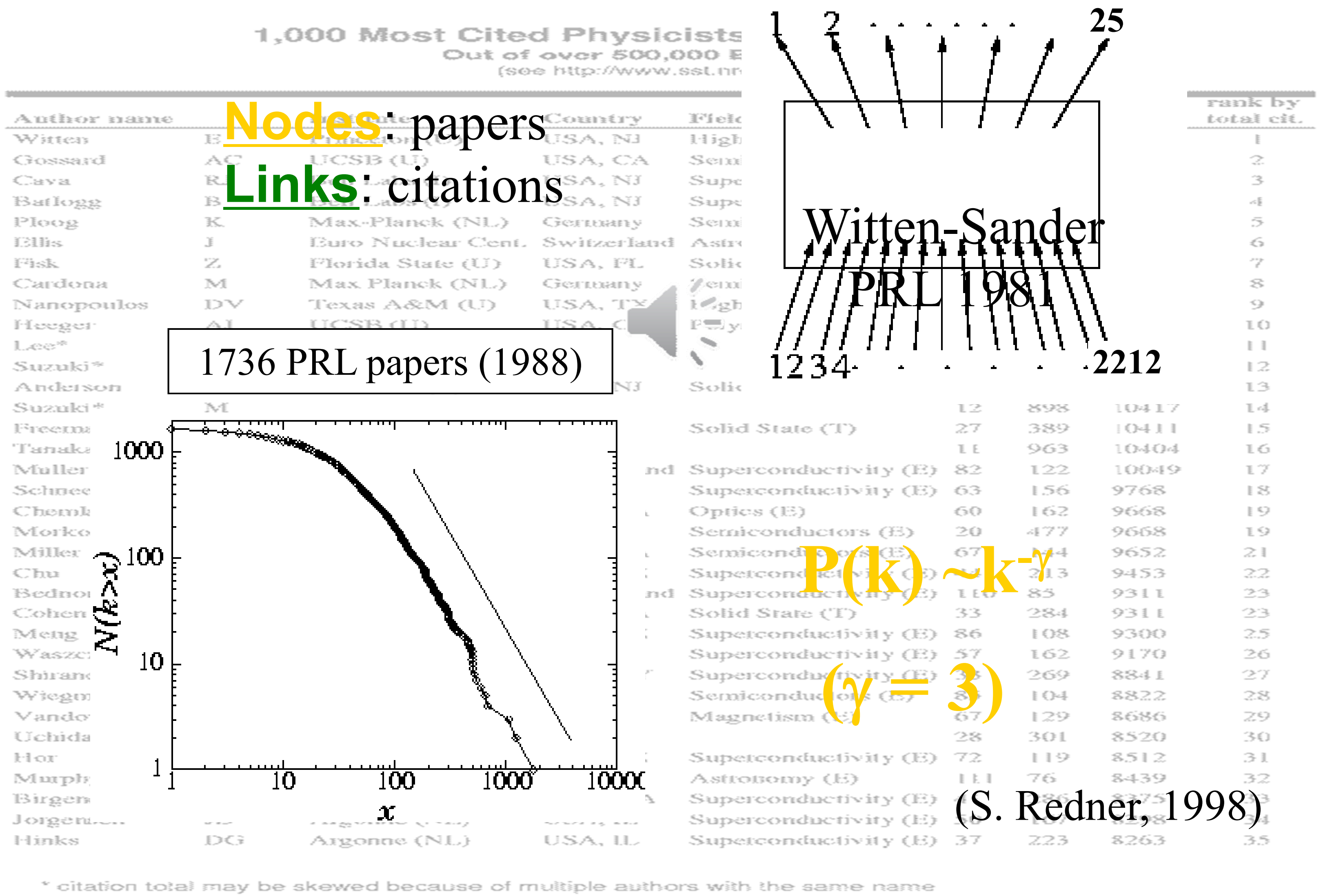
$P(k) \sim k^{-\gamma}$

$\gamma = 2.3$

Nodes: actors
Links: cast jointly

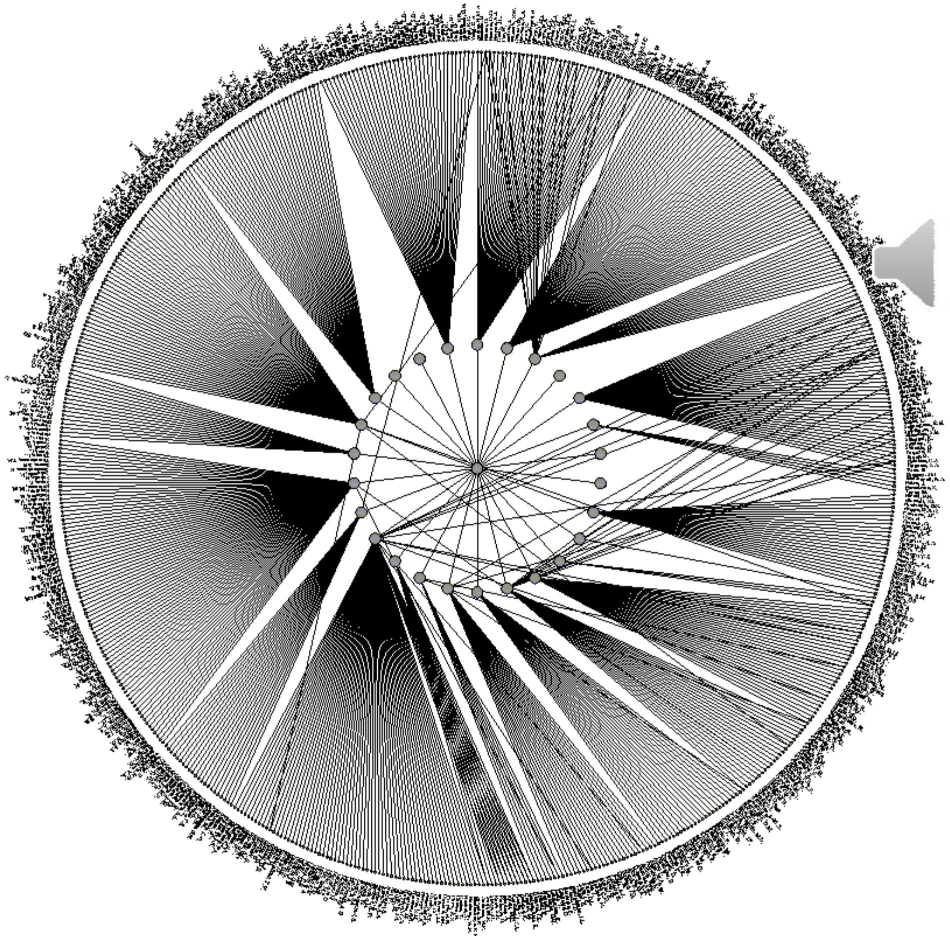


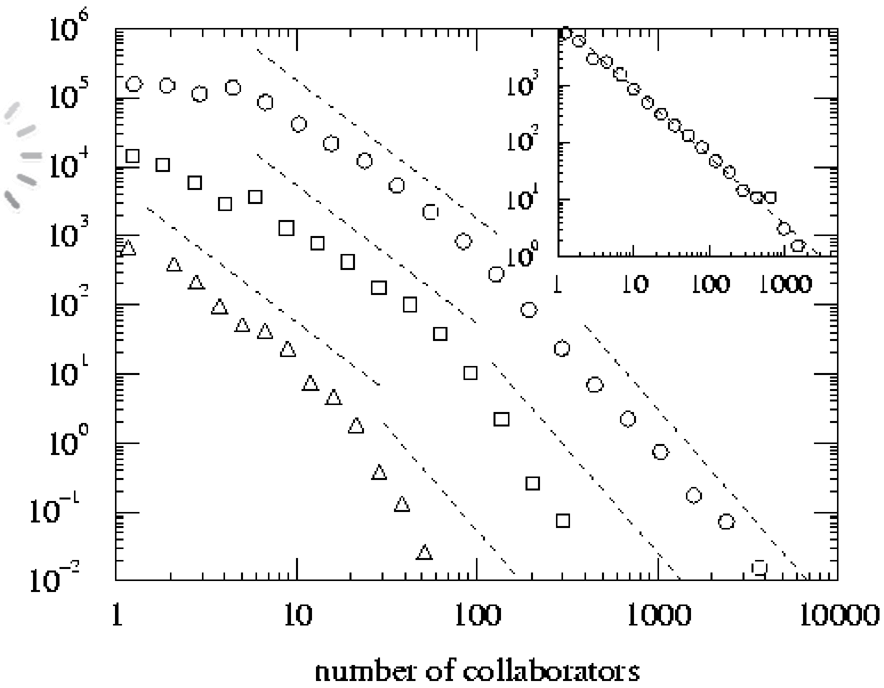
Case 3: Science Citation Index



Case 4: Science Coauthorship

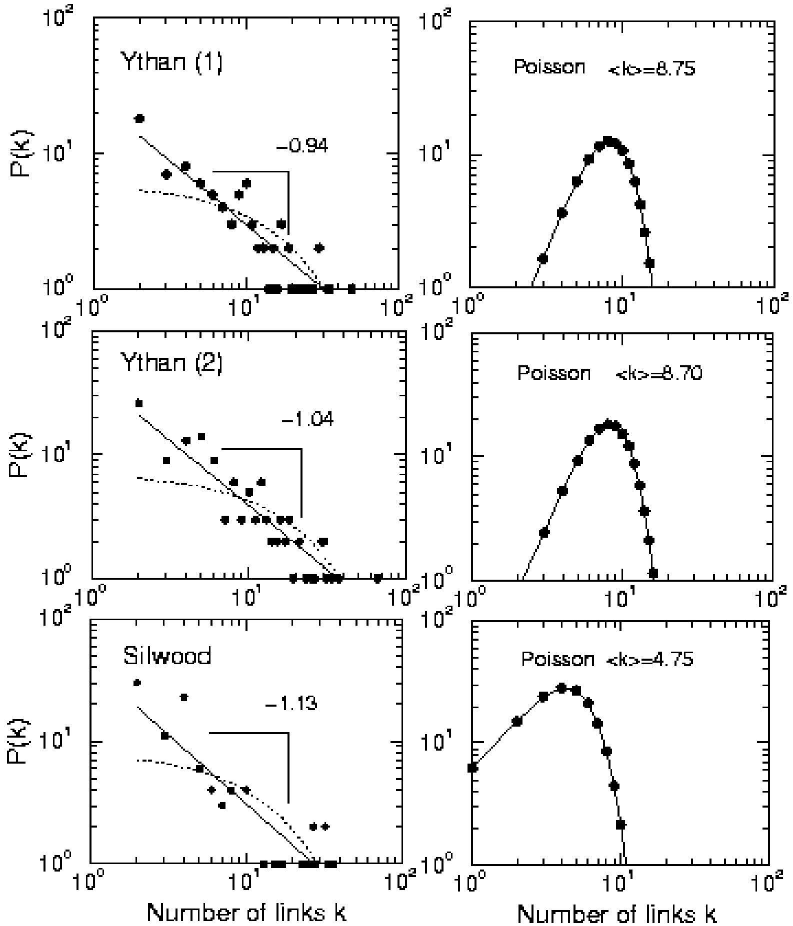
Nodes: scientist (authors)
Links: write paper together





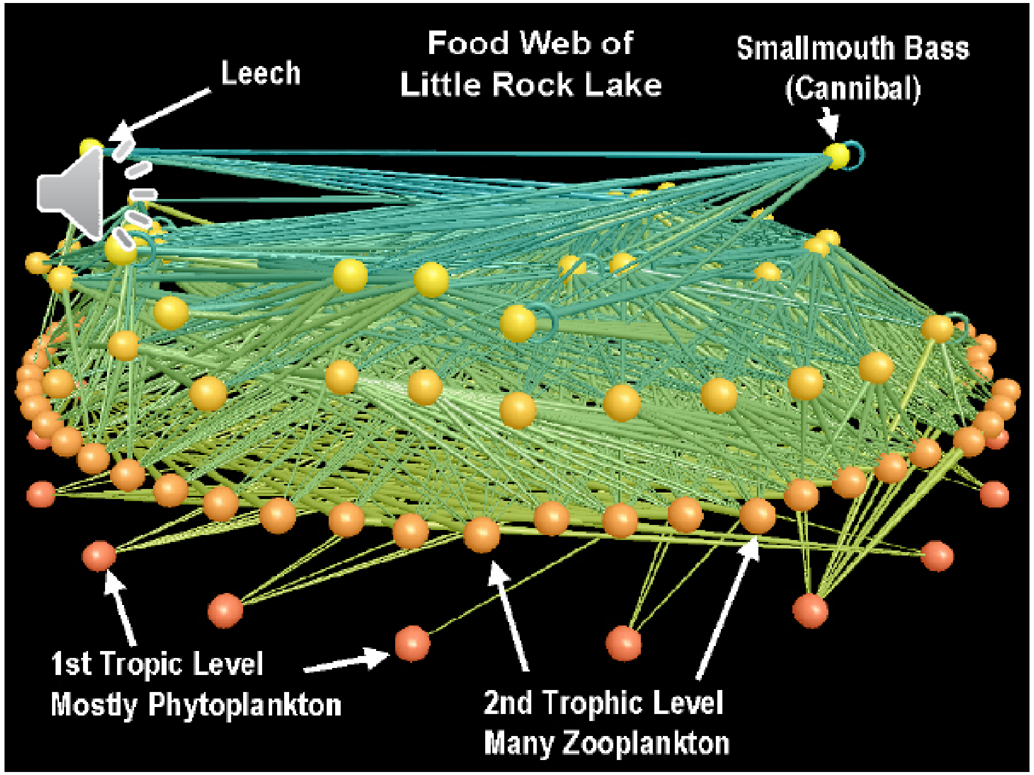
(Newman, 2000, H. Jeong et al 2001)

Case 5: Food Web



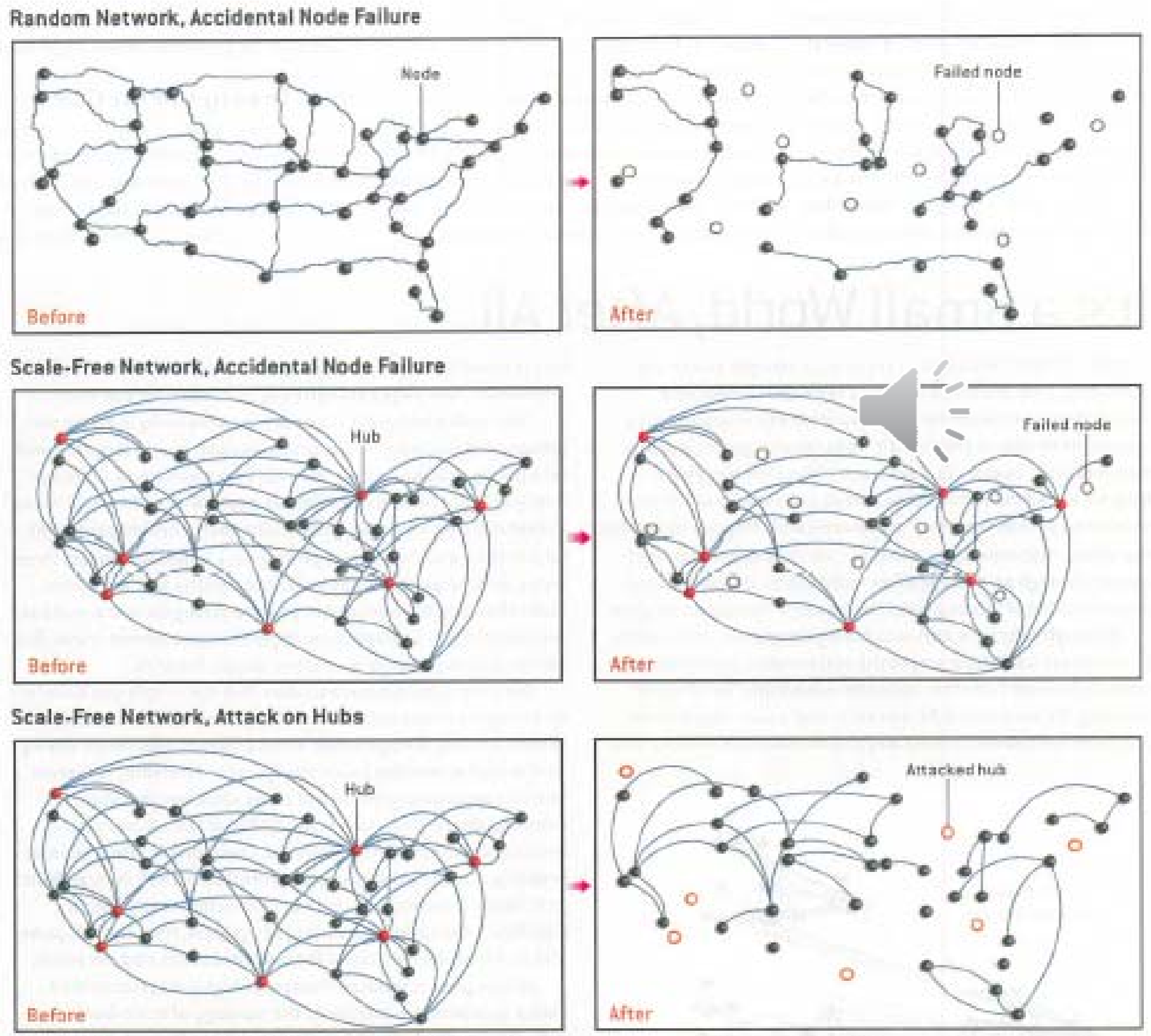
R. Sole (cond-mat/0011195)

Nodes: trophic species
Links: trophic interactions



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

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.

Thanks!

- Jiwon Hong (nowiz@hanyang.ac.kr)