Introduction to Networks

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

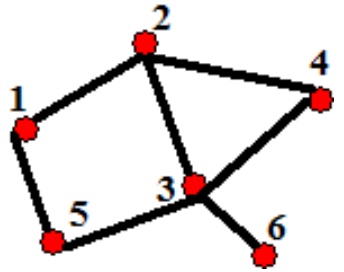
JANUARY 26, 2017

Introduction to Networks

- Basic Measures of Networks 
- Centrality Analysis in Networks
- Modeling of Network Formation
- Community Discovery and Cluster Analysis
- A Brief Introduction to Social Networks
- Summary

Networks and Their Representations

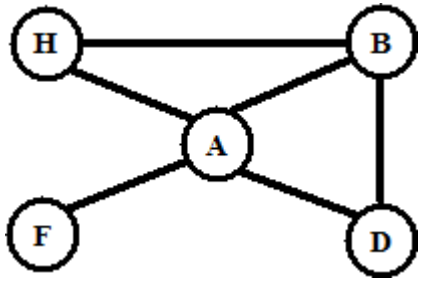
- A network/graph: G = (V, E), where V: vertices/nodes, E: edges/links
  - E: a subset of V × V, n = |V| (order of G), m = |E| (size of G)
  - Adjacency matrix
    - A_{ij} = 1 if there is an edge between vertices i and j; 0 otherwise
- Various kinds of networks:
  - Simple network: If a network has neither self-edges nor multi-edges
  - Multi-edge (multigraph): if more than one edge between the same pair of vertices
  - Self-loop: if an edge connects vertex to itself (i.e., (v_i, v_i))
  - Directed graph (digraph): if each edge has a direction (tail → head)
    - A_{ij} = 1 if there is an edge from j to i; 0 otherwise
  - Weighted graph: If a weight w_{ij} (a real number) is associated with each edge v_{ij}



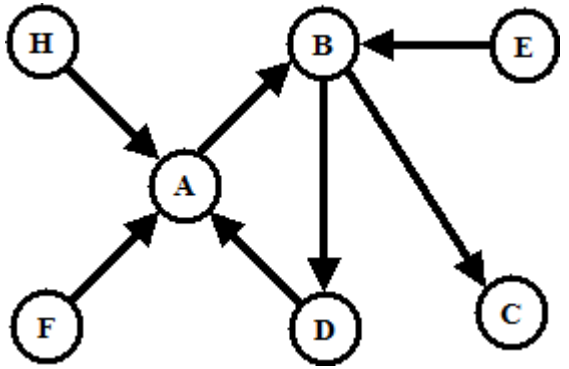
$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Vertex Degree for Undirected & Directed Networks

- Let a network G = (V, E)
- Undirected Network $d(v_i) = |v_j| \ s.t. \ e_{ij} \in E \wedge e_{ij} = e_{ji}$
  - Degree (or degree centrality) of a vertex: $d(v_i)$
  - # of edges connected to it, e.g., d(A) = 4, d(H) = 2



- Directed network
  - In-degree of a vertex $d_{in}(v_i)$: $d_{in}(v_i) = |v_j| \ s.t. \ e_{ij} \in E$
    - # of edges pointing to v_i
    - E.g., $d_{in}(A) = 3$, $d_{in}(B) = 2$
  - Out-degree of a vertex $d_{out}(v_i)$:
    - # of edges from v_i $d_{out}(v_i) = |v_j| \ s.t. \ e_{ji} \in E$
    - E.g., $d_{out}(A) = 1$, $d_{out}(B) = 2$



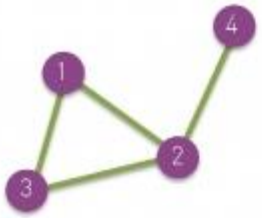
Various Kinds of Networks and Their Average Degrees

- Undirected
- Self-loops
- Multi-graph
- Directed
- Weighted
- Complete graph

L: toal # of links
<*k*>: average degree

Slides from L. Barabasi, Network Science, 2016

a. Undirected

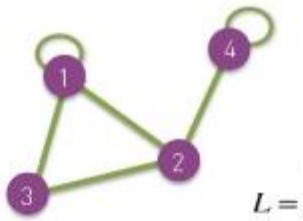


$$A_{ij}=\begin{pmatrix}0&1&1&0\\1&0&1&1\\1&1&0&0\\0&1&0&0\end{pmatrix}$$

$A_{ii}=0$ $\quad A_{ij}=A_{ji}$

$L=\frac{1}{2}\sum_{i,j=1}^{N}A_{ij}$ $\quad <k>=\frac{2L}{N}$

b. Self-loops

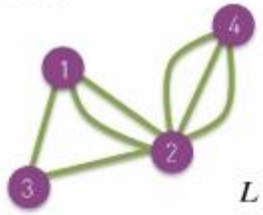


$$A_{ij}=\begin{pmatrix}1&1&1&0\\1&0&1&1\\1&1&0&0\\0&1&0&1\end{pmatrix}$$

$\exists i, A_{ii}\neq 0$ $\quad A_{ij}=A_{ji}$

$L=\frac{1}{2}\sum_{i,j=1,i\neq j}^{N}A_{ij}+\sum_{i=1}^{N}A_{ii}$ $\quad ?$

c. Multigraph (undirected)

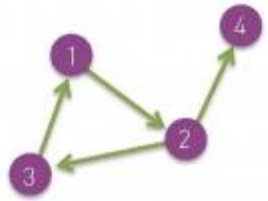


$$A_{ij}=\begin{pmatrix}0&2&1&0\\2&0&1&3\\1&1&0&0\\0&3&0&0\end{pmatrix}$$

$A_{ii}=0$ $\quad A_{ij}=A_{ji}$

$L=\frac{1}{2}\sum_{i,j=1}^{N}A_{ij}$ $\quad <k>=\frac{2L}{N}$

d. Directed

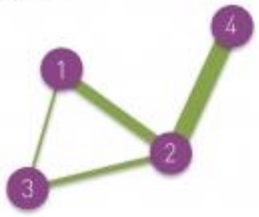


$$A_{ij}=\begin{pmatrix}0&1&0&0\\0&0&1&1\\1&0&0&0\\0&0&0&0\end{pmatrix}$$

$A_{ij}\neq A_{ji}$

$L=\sum_{i,j=1}^{N}A_{ij}$ $\quad <k>=\frac{L}{N}$

e. Weighted (undirected)

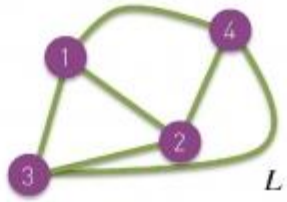


$$A_{ij}=\begin{pmatrix}0&2&0.5&0\\2&0&1&4\\0.5&1&0&0\\0&4&0&0\end{pmatrix}$$

$A_{ii}=0$ $\quad A_{ij}=A_{ji}$

$<k>=\frac{2L}{N}$

f. Complete Graph (undirected)



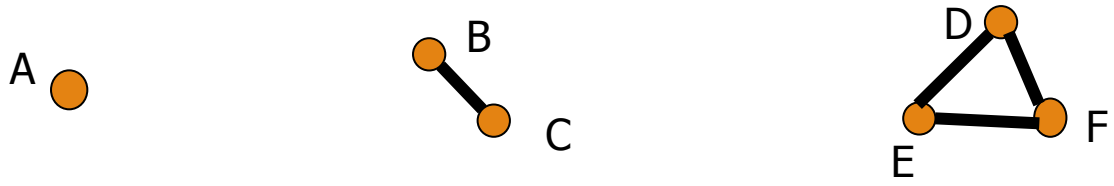
$$A_{ij}=\begin{pmatrix}0&1&1&1\\1&0&1&1\\1&1&0&1\\1&1&1&0\end{pmatrix}$$

$A_{ii}=0$ $\quad A_{i\neq j}=1$

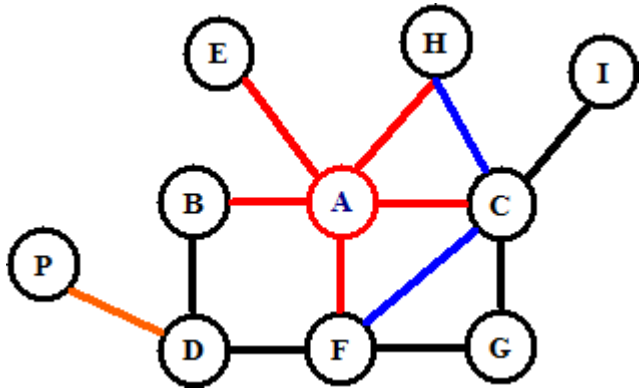
$L=L_{\max}=\frac{N(N-1)}{2}$ $\quad <k>=N-1$

Basic Network Structures and Properties

- **Subgraph**: A subset of the nodes and edges in a graph/network
  - Given a subset of vertices $V' \subseteq V$, the **induced subgraph** $G' = (V', E')$ consists exactly of all the edges present in G between vertices in V'
- **Clique** (complete graph): Every node is connected to every other
- **Singleton** vs. **dyad** (two nodes and their relationship) vs. **triad**:



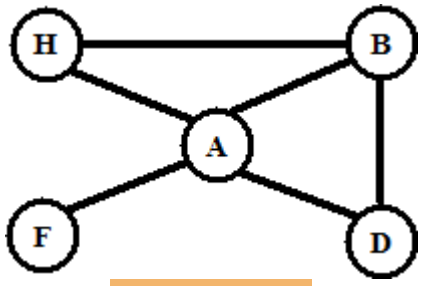
- **Ego-centric network**: A network pull out by selecting a node and all of its connections
  - The 1-degree egocentric network of A
  - The 1.5-degree egocentric network of A
  - The 2-degree egocentric network of A



Degree Distribution and Path

- **Degree sequence** of a graph: The list of degrees of the nodes sorted in non-increasing order
  - E.g., in graph G_1, degree sequence: (4, 3, 2, 2, 1)
- **Degree frequency distribution** of a graph: Let N_k denote the # of vertices with degree k
  - (N_0, N_1, …, N_t), t is max degree for a node in G
  - E.g., in graph G_1, degree frequency distribution: (0, 1, 2, 1, 1)
- **Degree distribution** of a graph:

  Probability mass function f for random variable X
  - (f(0), f(1), …, f(t), where $f(k) = P(X = k) = N_k/n$
  - E.g., in graph G_1, degree distrib.: (0, 0.2, 0.4, 0.2, 0.2)



Graph G_1

- **Walk** in a graph G between nodes X and Y: ordered sequence of vertices, starting at X and ending at Y, s.t. there is an edge between every pair of consecutive vertices
  - **Hops**: the length of the walk
- **Path**: a walk with distinct vertices
  - **Distance**: the length of the shortest path

Paths

- Path: A sequence of vertices that every consecutive pair of vertices in the sequence is connected by an edge in the network
- Length of a path: # of edges traversed along the path
- Total # of path of length 2 from *j* to *i*, via any vertex in $N_{ij}^{(2)}$ is

$$N_{ij}^{(2)} = \sum_{k=1}^{n} A_{ik} A_{kj} = [A^2]_{ij}$$

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

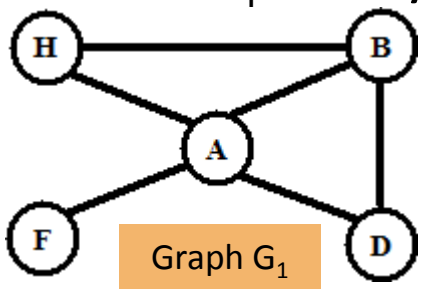
- Generalizing to path of arbitrary length, we have: $N_{ij}^{(r)} = [A^r]_{ij}$
- When starting and ending at the same vertex *i*, we have: $L_r = \sum_{i=1}^{n} [A^r]_{ii} = Tr\, A^r$
- # of loops can be expressed in terms of adjacency matrix
  - Matrix A written in the form of $A = UKU^T$, where U is the orthogonal matrix of eigenvectors and K is the diagonal matrix of eigenvalues

$$L_r = Tr\,(UKU^T)^r = Tr\,(UK^rU^T) = Tr\,(UU^TK^r) = Tr\,(K^r) = \sum_i k_i^r$$

    - where k_i is the *i*-th eigenvalue of the adjacency matrix

Radius and Diameter of a Network

- **Eccentricity**: The eccentricity of a node v_i is the maximum distance from v_i to any other nodes in the graph
  - $e(v_i) = \max_j \{d(v_i, v_j)\}$
  - E.g., e(A) = 1, e(F) = e(B) = e(D) = e(H) = 2

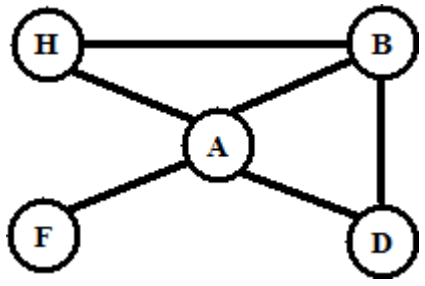


Graph G_1

- **Radius** of a connected graph G: the min eccentricity of any node in G
  - $r(G) = \min_i \{e(v_i)\} = \min_i \{\max_j \{d(v_i, v_j)\}\}$
  - E.g., $r(G_1) = 1$
- **Diameter** of a connected graph G: the max eccentricity of any node in G
  - $d(G) = \max_i \{e(v_i)\} = \max_{i,j} \{d(v_i, v_j)\}$
  - E.g., $d(G_1) = 2$
- Diameter is sensitive to outliers. Effective diameter: min # of hops for which a large fraction, typically 90%, of all connected pairs of nodes can reach each other

Various Kinds of Paths

- Shortest path (geodesic path, *d*):
  - Geodesic paths are not necessarily unique: It is quite possible to have more than one path of equal length between a given pair of vertices
  - *Diameter* of a graph: the length of the longest geodesic path between any pair of vertices in the network for which a path actually exists

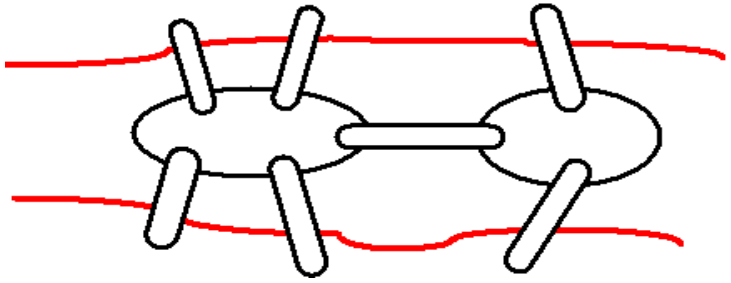


For this graph, what is <d> ?

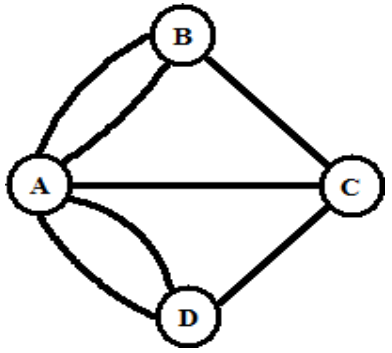
- Average path length (<d>):
  - Average of the shortest paths between all pairs of nodes

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j=1,N(i \neq j)} d_{i,j}$$

- Eulerian path: a path that traverses each edge in a network exactly once



The Königsberg bridge problem

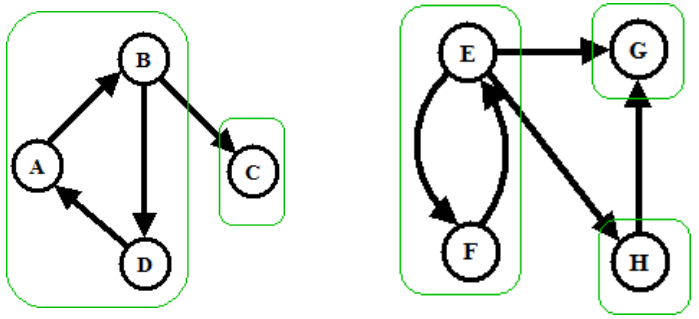


- Hamilton path: A path that visits each vertex in a network exactly once

Components in Directed & Undirected Network

- Undirected network:
  - Component: A subset of the vertices of a network such that there exists at least one path from vertex to each other vertex
  - Adjacency matrix of a network with more than one component can be written in block diagraph form

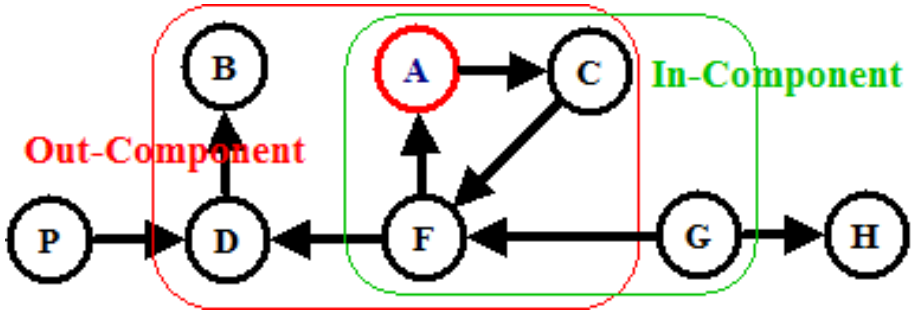
$$A = \begin{pmatrix} [\ ] & 0 & \cdots \\ 0 & [\ ] & \cdots \\ \vdots & \cdots & \ddots \end{pmatrix}$$



multiple size-2 edge/vertex cut set

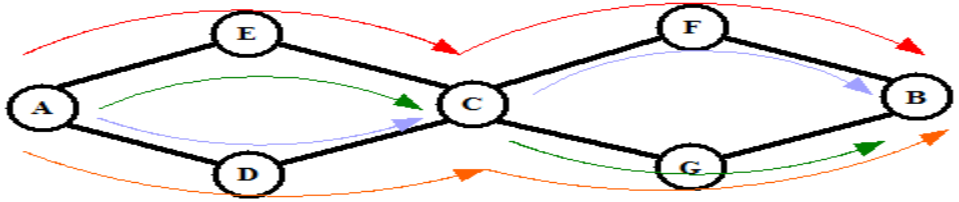
The graph contains 2 weakly connected and 5 strongly connected components

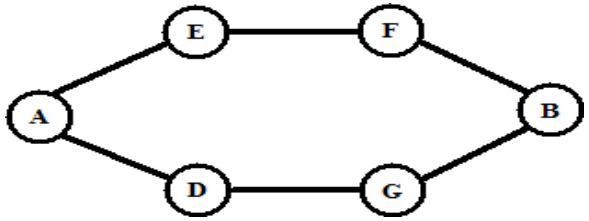
- Directed network:
  - Weakly vs. strongly connected components
    - Weakly connected: if the vertices are connected by 1 or more paths when one can go either way along any edge
    - Out-component vs. in-component of a vertex (A) or a strongly connected component (A-C-F)
    - Out-component: Those reachable from vertex A



Independent Paths, Connectivity, and Cut Sets

- Two path connecting a pair of vertices (A, B) are **edge-independent** if they share no edges
- Two path are **vertex-independent** if they share no vertices other than the starting and ending vertices



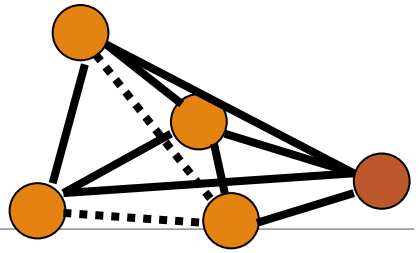


multiple size-2 edge/vertex cut set

- A **vertex cut set** is a set of vertices whose removal will disconnected a specified pair of vertices
- An **edge cut set** is a set of edges whose removal will disconnected a specified pair of vertices
- A **minimum cut set**: the smallest cut set that will disconnect a specified pair of vertices
- Menger's theorem => maxflow/min-cut theorem: For a pair of vertices,

  size of min-cut set = vertex connectivity = maximum flow
- This works also for weighted networks

Clustering Coefficient



- Real networks are sparse: Corresponding to a complete graph
- Clustering coefficient of a node v_i: A measure of the density of edges in the neighborhood of v_i
- Let $G_i = (V_i, E_i)$ be the subgraph induced by the neighbors of vertex v_i, $|V_i| = n_i$ (# of neighbors of v_i), and $|E_i| = m_i$ (# of edges among the neighbors of v_i)
- **Clustering coefficient** of v_i for **undirected network** is

$$C(v_i) = \frac{\#\ edges\ in\ G_i}{max\ \#\ edges\ in\ G_i} = \frac{m_i}{\binom{n_i}{2}} = \frac{2 \times m_i}{n_i(n_i-1)}$$

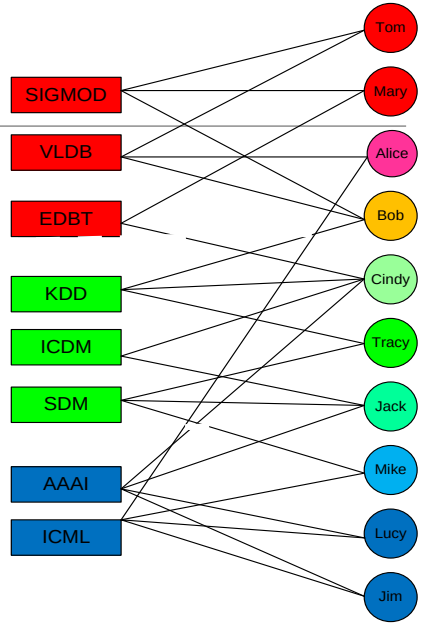
(corresp. to when G_i is a complete graph)

  - For directed network,

$$C(v_i) = \frac{\#\ edges\ in\ G_i}{max\ \#\ edges\ in\ G_i} = \frac{m_i}{n_i(n_i-1)}$$

- Clustering coefficient of a graph G: $C(G) = \frac{1}{n}\sum_i C(v_i)$
  - Averaging the local clustering coefficient of all the vertices (Watts & Strogatz)

Bipartite Networks



- Bipartite Network: two kinds of vertices, and edges linking only vertices of unlike types
- Incidence matrix:
  - B_{ij} = 1 if vertex j links to group i

    0 otherwise
- One can create a one-mode project from the two-mode partite form (but with info loss)
- The projection to one-mode can be written in terms of the incidence matrix B as follows

$$P_{ij} = \sum_{k=1}^{g} B_{ki}\, B_{kj} = \sum_{k=1}^{g} B_{ik}{}^{T}\, B_{kj}$$

- The product of $B_{ki}\, B_{kj}$ will be 1 if *i* and *j* both belong to the sample group *k* in the bi-partite network

Co-citation and Bibliographic Coupling

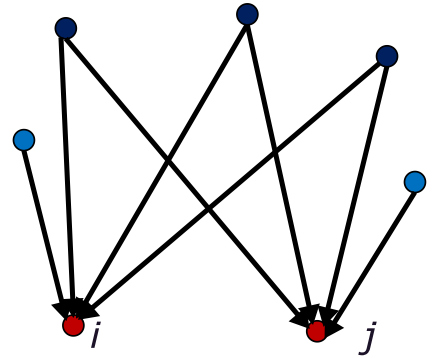
- Co-citation of vertices i and j: $A_{ik}A_{jk} = 1$ if i and j are both cited by k
  - # of vertices having outgoing edges pointing to both i and j

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

  - Co-citation 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

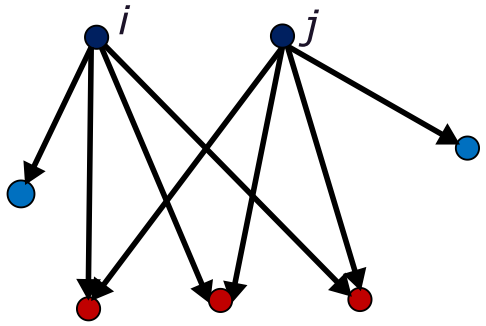
- Bibliographic coupling of vertices i and j:

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

  - Bibliographic coupling of i and j:

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

  - Bibliographic coupling 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

Co-citation & Bibliographic Coupling: A Comparison

- Two measures are affected by the number of incoming and outgoing edges that vertices have
- For strong co-citation: 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 co-citation and bibliographic coupling

Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks 
- Modeling of Network Formation
- Primitives of Social Networks
- Summary

Centrality: Basic Measure in a Network

- Centrality: How "central" a node is in the network
- **Degree centrality**: degree of a node (the higher degree, more important the node)
- **Eccentricity centrality**: the less eccentric, the more central
  - $c(v_i) = 1/e(v_i)$
  - Central node: $e(v_i) = r(G)$ (if it equals the radius of G)
  - Periphery node: $e(v_i) = d(G)$ (if it equals the diameter of G)
  - Often used in facility location, e.g., emergency center
- **Closeness centrality**: the average of the shortest path length from the node to every other node in the network, indicating how close a node is to all other nodes in the network
  - $c(v_i) = 1/\sum_j d(v_i, v_j)$
  - median node v_m if v_m has the smallest total distance $\sum_j d(v_m, v_j)$
  - Facility location, e.g., shopping center, minimize total distance

Centrality Measures (II)

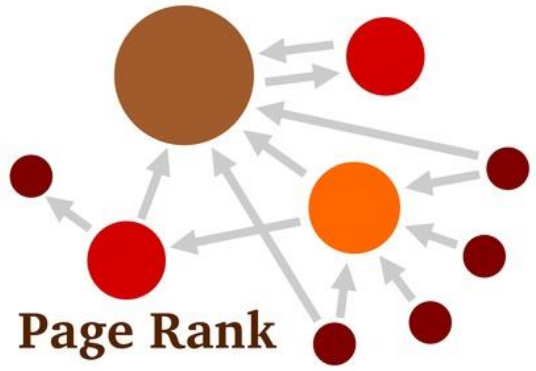
- **Betweenness centrality** for a node v: # of shortest paths from all vertices to all others that pass through v
  - η_{jk}: # of shortest paths between vertices v_j and v_k
  - $\eta_{jk}(v_i)$: # of such paths that contain v_i
  - Betweenness centrality of a vertex v_i:

$$c(v_i) = \sum_{j \neq i} \sum_{k \neq i, k > j} \frac{\eta_{jk}(v_i)}{\eta_{jk}}$$

  - Indicating a central "monitoring role" played by v_i for various pairs of nodes
- **Eigenvector centrality**: Measure the influence of a node in a network, i.e., connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes

Centrality Measures on the Web (I): Eigenvector Centrality

- Web: a directed graph, PageRank and HITS are typical algorithms
- **Eigenvector centrality**, or **prestige**, importance, or rank of a node v
  - The more nodes point to v, the higher v's prestige
  - The more prestige of a node pointing to v, the higher v's prestige
- Let p(u) be prestige score for node u. Then

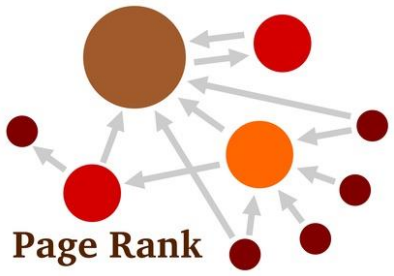


$$p(v) = \sum_u A(u,v) \cdot p(u) = \sum_u A^T(v,u) \cdot p(u)$$

- Written in vector form: $\mathbf{p}' = \mathbf{A}^T\mathbf{p}$
- At the k-th iteration, we have $\mathbf{p}_k = (\mathbf{A}^T)^k\mathbf{p}_0$
  - Vetor $\mathbf{p}_k$ converges to the dominant eigenvector of $\mathbf{A}^T$ with increasing k

Centrality Measures on the Web (II): PageRank

- **Random surfing assumption:** A web surfer randomly chooses one of the outgoing links from the current page or with some very small probability randomly jumps to any other page in the web graph



- *Pagerank* of a page v: the probability of a random web surfer landing at v
- **Normalized prestige**:
  - The prob. of visiting a page pointed by v is $1/d_{out}(v)$, d_{out} is outdegree of v
  - Compute updated pagerank vector for v,

$$p(v) = \sum_u \frac{\mathbf{A}(u,v)}{d_{out}}(u) \cdot p(u) = \sum_u \mathbf{N}(u,v) \cdot p(u) = \sum_u \mathbf{N}^T(v,u) \cdot p(u), \ or \ \ \mathbf{p} = \mathbf{N}^T \cdot \mathbf{p}$$

where N(u, v) is the normalized adjacency matrix of the graph, and

N(u, v) = $1/d_{out}(u)$ if (u, v) in E or 0 o.w.

- **Random Jumps:** a small prob. jumping to any other node (viewing web as a fully connected graph, i.e., adjacency matrix $\mathbf{A}_r = \mathbf{1}_{nXn}$)

$$p(v) = \sum_u \frac{\mathbf{A}_r(u,v)}{d_{out}}(u) \cdot p(u) = \sum_u \mathbf{N}_r(u,v) \cdot p(u) = \sum_u \mathbf{N}_r{}^T(v,u) \cdot p(u), \ or \ \ \mathbf{p} = \mathbf{N}_r{}^T \cdot \mathbf{p}$$

- Pagerank score computation:

$$\mathbf{p}' = (1-\alpha)\mathbf{N}^T\mathbf{p} + \alpha\mathbf{N}_r{}^T\mathbf{p} = ((1-\alpha)\mathbf{N}^T + \alpha\mathbf{N}_r{}^T)\mathbf{p} = \mathbf{M}^T\mathbf{p}$$

PageRank: Capturing Page Popularity (Brin & Page'98)

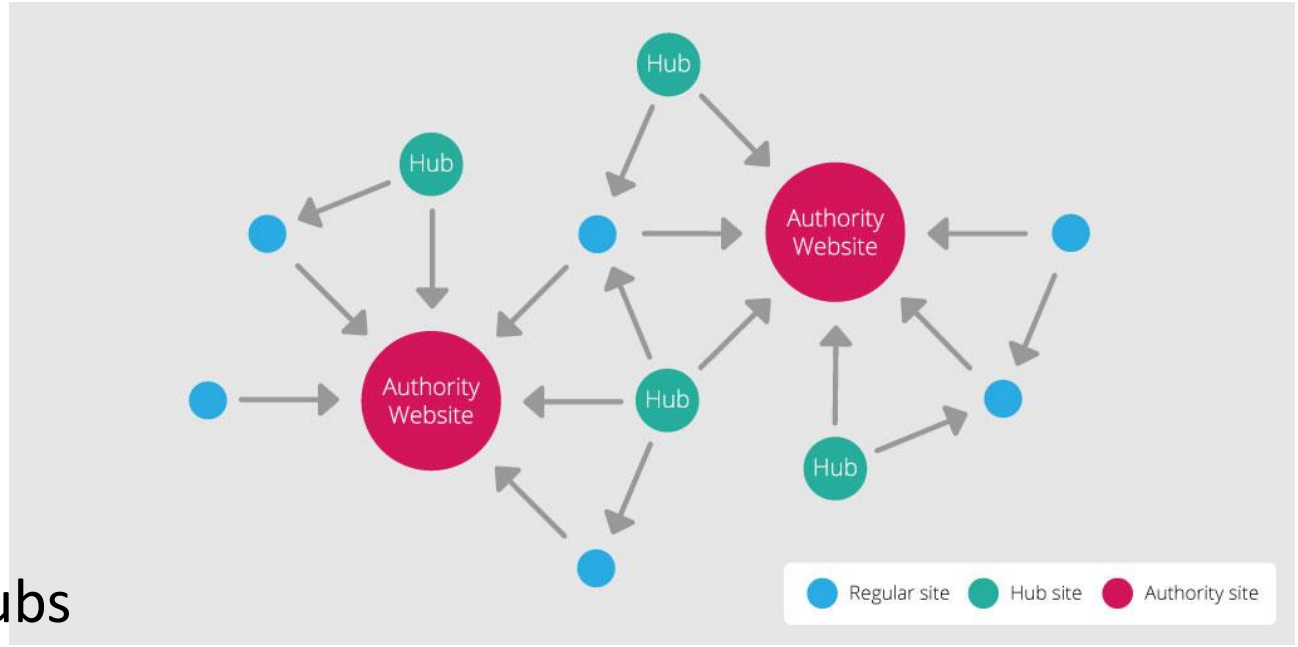
- Intuitions
  - Links are like citations in literature
  - A page that is cited often can be expected to be more useful in general
- PageRank is essentially "citation counting", but improves over simple counting
  - Consider "indirect citations" (being cited by a highly cited paper counts a lot…)
  - Smoothing of citations (every page is assumed to have a non-zero citation count)
- PageRank can also be interpreted as a random surfing model (thus capturing popularity)
  - At any page,

    With prob. α, randomly jumping to a page

    With prob. $(1 - \alpha)$, randomly picking a link to follow

HITS: Capturing Authorities & Hubs (Kleinberg'98)

- Intuitions of HITS (Hyperlink Induced Topic Search)
  - Pages that are widely cited are good authorities
  - Pages that cite many other pages are good hubs
- The key idea of HITS
  - Good authorities are cited by good hubs
  - Good hubs point to good authorities
  - Iterative reinforcement …
- $\mathbf{AA}^T$ is the co-citation matrix and $\mathbf{A}^T\mathbf{A}$ is the bibliographic coupling matrix. Authority centrality is eigenvector centrality for the co-citation network



Centrality Measures on the Web (III): HITS (Computing Hub & Authority Scores)

- For a specific query, a page of high Pagerank score may not be that relevant
- HITS (Hyperlink Induced Topic Search) computes two values for a page
  - Authority score: analogous to pagerank/prestige scores
  - Hub score: based on how many "good" pages it points to
- How is HITS query-based?
  - first uses standard search engines to retrieve the set of relevant pages
  - then expands the set to include any page that point to or is pointed to by) some pages in the set
  - Any pages originating from the same host are eliminated
  - HITS is only applied on this expanded query-specific graph G
- Computation: $a(v) = \sum_u \mathbf{A}^T(v,u) \cdot h(u)$ $\qquad h(v) = \sum_u \mathbf{A}(v,u) \cdot a(u)$
- In matrix computation (essentially two eigenvector computation):

$$\mathbf{a}_k = \mathbf{A}^T\mathbf{h}_{k-1} = \mathbf{A}^T(\mathbf{A}\mathbf{a}_{k-2}) = (\mathbf{A}^T\mathbf{A})\mathbf{a}_{k-2} \qquad \mathbf{h}_k = \mathbf{A}\mathbf{a}_{k-1} = \mathbf{A}(\mathbf{A}^T\mathbf{h}_{k-2}) = (\mathbf{A}\mathbf{A}^T)\mathbf{h}_{k-2}$$

Metrics (Measures) in Social Network Analysis (I)

- **Betweenness**: The extent to which a node lies between other nodes in the network. This measure takes into account the connectivity of the node's neighbors, giving a higher value for nodes which bridge clusters. The measure reflects the number of people who a person is connecting indirectly through their direct links
- **Bridge**: An edge is a bridge if deleting it would cause its endpoints to lie in different components of a graph
- **Centrality**: This measure gives a rough indication of the social power of a node based on how well they "connect" the network. "Betweenness", "Closeness", and "Degree" are all measures of centrality
- **Centralization**: The difference between the number of links for each node divided by maximum possible sum of differences. A centralized network will have many of its links dispersed around one or a few nodes, while a decentralized network is one in which there is little variation between the number of links each node possesses
- **Closeness**: The degree an individual is near all other individuals in a network (directly or indirectly). It reflects the ability to access information through the "grapevine" of network members. Thus, closeness is the inverse of the sum of the shortest distances between each individual and every other person in the network

Metrics (Measures) in Social Network Analysis (II)

- **Clustering coefficient**: A measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'
- **Cohesion**: The degree to which actors are connected directly to each other by *cohesive* bonds. Groups are identified as '*cliques*' if every individual is directly tied to every other individual, '*social circles*' if there is less stringency of direct contact, which is imprecise, or as *structurally cohesive* blocks if precision is wanted
- **Degree (or geodesic distance):** The count of the number of ties to other actors in the network
- **(Individual-level) Density:** The degree a respondent's ties know one another/ proportion of ties among an individual's nominees. Network or global-level density is the proportion of ties in a network relative to the total number possible (sparse versus dense networks)
- **Flow betweenness centrality**: The degree that a node contributes to sum of maximum flow between all pairs of nodes (not that node)
- **Eigenvector centrality**: A measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to nodes having a high score contribute more to the score of the node in question

Metrics (Measures) in Social Network Analysis (III)

- **Local Bridge**: An edge is a local bridge if its endpoints share no common neighbors. Unlike a bridge, a local bridge is contained in a cycle
- **Path Length**: The distances between pairs of nodes in the network. Average path-length is the average of these distances between all pairs of nodes
- **Prestige**: In a directed graph prestige is the term used to describe a node's centrality. "Degree Prestige", "Proximity Prestige", and "Status Prestige" are all measures of Prestige
- **Radiality Degree**: an individual's network reaches out into the network and provides novel information and influence
- **Reach**: The degree any member of a network can reach other members of the network
- **Structural cohesion**: The minimum number of members who, if removed from a group, would disconnect the group
- **Structural equivalence**: Refers to the extent to which nodes have a common set of linkages to other nodes in the system. The nodes don't need to have any ties to each other to be structurally equivalent.
- **Structural hole**: Static holes that can be strategically filled by connecting one or more links to link together other points

Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks
- Modeling of Network Formation 
- Primitives of Social Networks
- Summary

Why Network Modeling?

- Many real-world networks exhibit certain common characteristics, even though they come from different domains, e.g., communication, social, and biological networks
- A typical network has the following common properties:
  - *Few* connected components:
    - often only 1 or a small number, independent of network size
  - *Small* diameter:
    - often a constant independent of network size (like 6)
    - 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
  - A *heavy-tailed* degree distribution:
    - a small but reliable number of high-degree vertices
    - often of *power law* form

Probabilistic Models of Networks

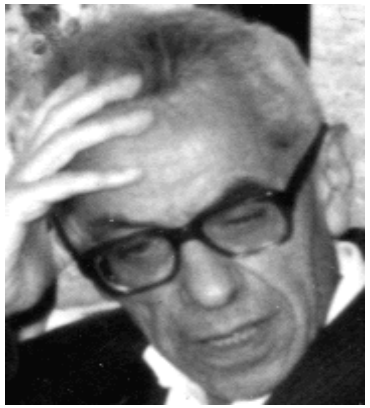
- All of the network generation models we will study are *probabilistic* or *statistical* in nature
- They can generate networks of any size
- They often have various *parameters* that can be set:
  - size of network generated
  - average degree of a vertex
  - fraction of long-distance connections
- The models generate a *distribution* over networks
- Statements are always *statistical* in nature:
  - *with high probability*, diameter is small
  - on average, degree distribution has heavy tail
- Thus, we're going to need some basic statistics and probability theory

Some Models of Network Generation

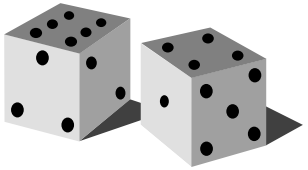
- ***Erdös-Rényi Random graph model:***
  - Gives few components and small diameter
  - does not give high clustering and heavy-tailed degree distributions
  - is the mathematically most well-studied and understood model
- ***Watts-Strogatz small world graph model:***
  - gives few components, small diameter and high clustering
  - does not give heavy-tailed degree distributions
- ***Barabási-Albert Scale-free model:***
  - gives few components, small diameter and heavy-tailed distribution
  - does not give high clustering
- *Hierarchical network:*
  - few components, small diameter, high clustering, heavy-tailed
- *Affiliation network:*
  - models group-actor formation

Erdös-Rényi (ER) 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*), in which **every possible edge occurs independently with probability *p***
  - *G*(*N, p*): a network of **N** nodes, each node pair is connected with probability of *p*
    - Paul Erdős and Alfréd Rényi: "On Random Graphs" (1959)
    - E. N. Gilbert: "Random Graphs" (1959) (proposed independently)
  - Usually, *N is large* and *p ~ 1/N*
    - Choices: p = 1/2N, p = 1/N, p = 2/N, p = 10/N, p = log(N)/N, etc.
- An alternative model: *G*(*N, L*), L: a fixed total # of edges, placed randomly



Pál Erdös (1913-1996)



Degree Distribution of Random Graphs

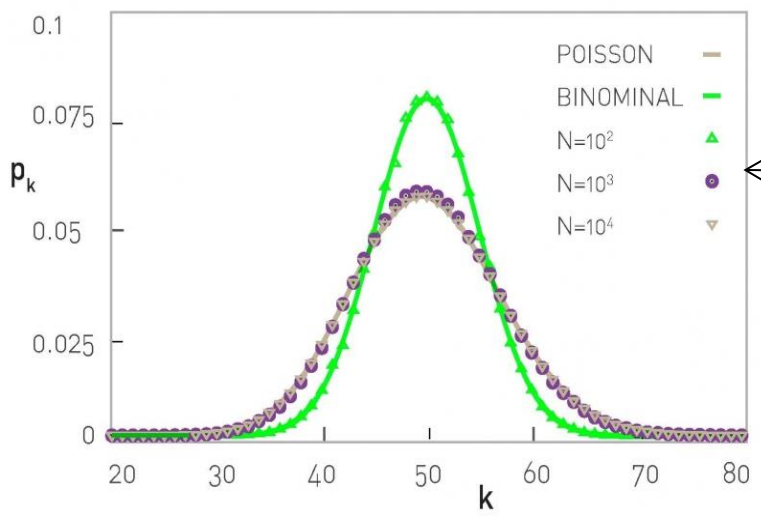
- The degree distribution of a random (small) network follows binomial distribution:
  - $$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

The probability that the remaining links are missing

The probability that k of its links are present

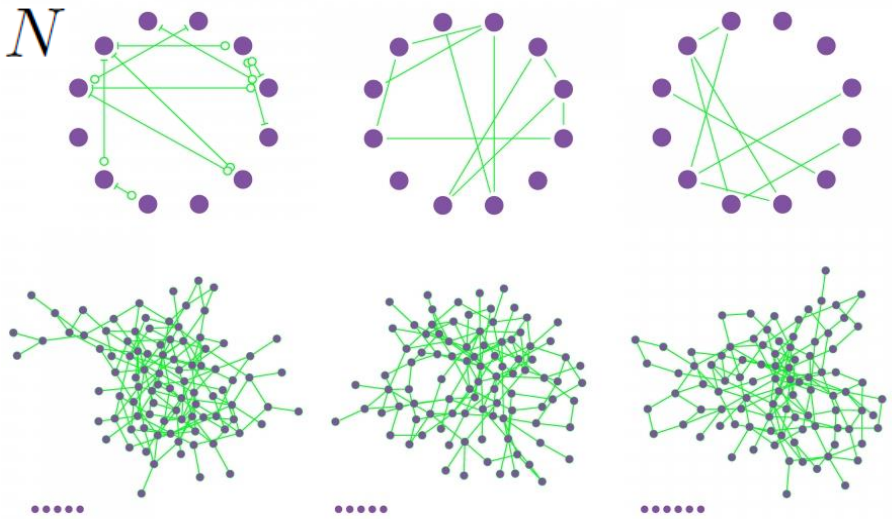
# of ways to select *k* links from *N–1* potential links

- Most real network are sparse, i.e., avg degree $\langle k \rangle \ll N$
  - The degree density of a large network is well approximated by Poisson distribution



$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

Degree distribution of a network with <k> = 50 and N= 10^2, 10^3, 10^4.



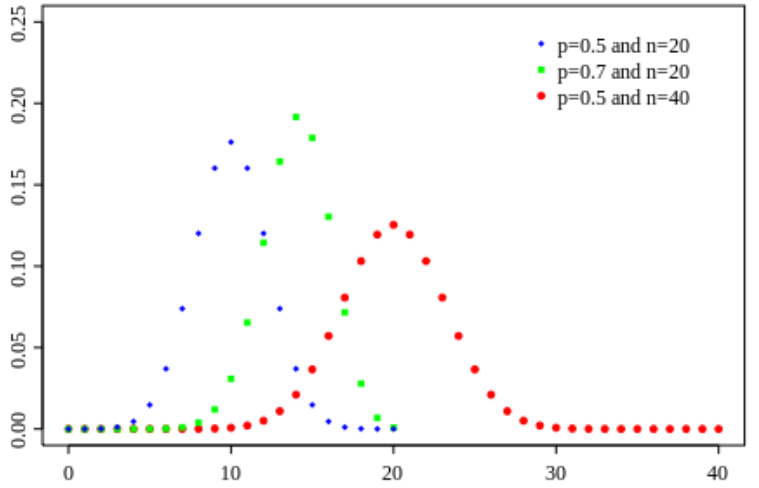
- Top row: 3 random networks generated with p=1/6 and N = 12
- Bottom: 3 random networks generated with p=0.03, N = 100

Figures courtesy Barabasi 2016

Review: Binomial Distribution vs. Poisson Distribution

- The Binomial distribution:
  - Coin with prob(heads) = *p*, toss *n* times, the probability of getting exactly *k* heads:
  - For large *n* and *fixed p:* approximated well by a normal distribution with $\mu = np$, $\sigma = sqrt(np(1-p))$

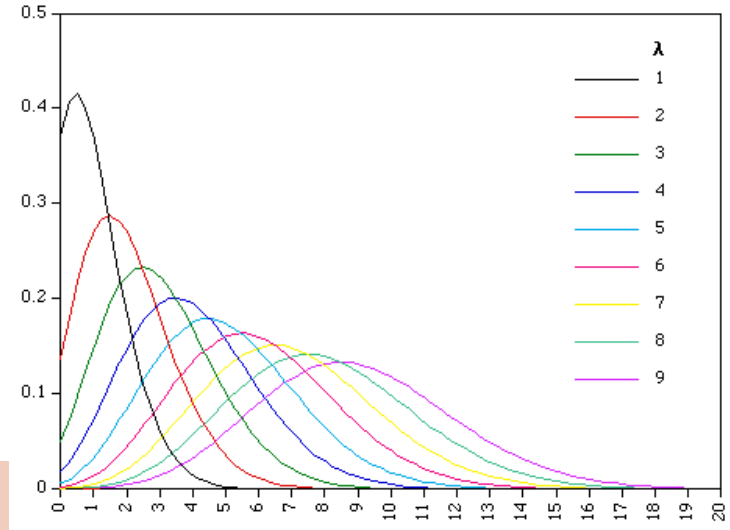
$$Pr\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$



- The Poisson distribution
  - An event can occur 0, 1, 2, … times in an interval
    - λ: The average # of events in an interval
  - We have:

$$Pr\{k \text{ events in interval}\} = \frac{\lambda^k e^{-\lambda}}{k!}$$

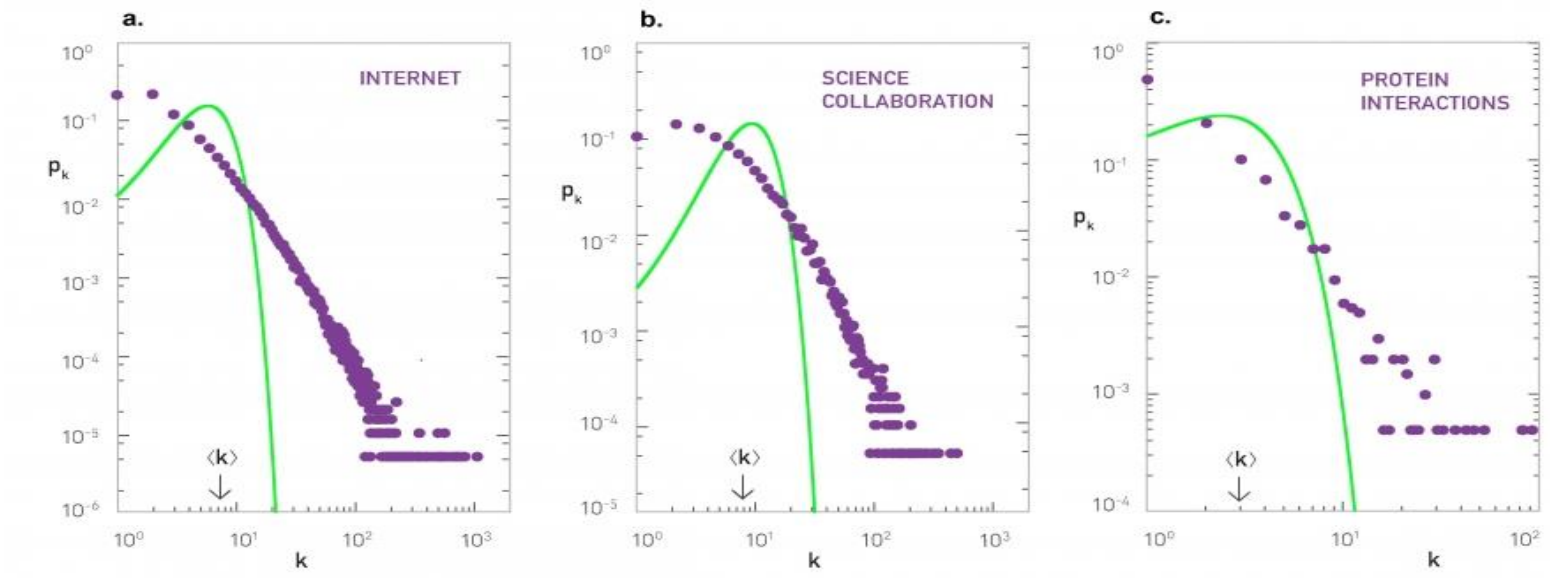
- Binomial distribution with large n, p = λ /n (λ fixed)
  - converges to Poisson distribution with mean λ



Figures courtesy Wikipedia

Real Networks Do Not Fit the Random Network Model

- Property of random networks: Comparable degrees
  - *In a large random network the degree of most nodes is in the narrow vicinity of ‹k›*
- Real networks are not Poisson (Real nets have hubs + wide degree spread)
  - The Poisson form significantly underestimates the number of high degree nodes
  - The spread in the degree of real networks is much wider than expected in a random network (if Internet were random, we expect $\sigma = 2.57$, but $\sigma_{Internet} = 14.14$)



The degree distribution of the (a) Internet, (b) science collaboration network, and (c) protein interaction network

Figures courtesy Barabasi 2016

The Small World Phenomenon & Erdös Number

- Small world phenomenon (Six degrees of separation)
  - Stanley Milgram's experiments (1960s)
  - Microsoft Instant Messaging (IM) experiment: J. Leskovec & E. Horvitz (WWW'08)
    - 240 M active user accounts: Est. avg. distance 6.6 & est. mean median 7
- Why small world? Let N: # of nodes, d: distance, <k>: average degree

$$N(d) \approx 1 + \langle k \rangle + \langle k \rangle^2 + \cdots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1} \approx \langle k \rangle^d$$

  - In today's tightly connected world: $d \approx \frac{\ln N}{\ln \langle k \rangle} \approx \frac{\ln(7 \times 10^9)}{\ln(10^3)} \approx 3.28$

World population

Average connectivity

- Erdös number: Distance from him/her to Erdös in the coauthor graph
  - Paul Erdös (a mathematician who published about 1500 papers)
  - Similarly, Kevin Bacon number (co-appearance in a movie)

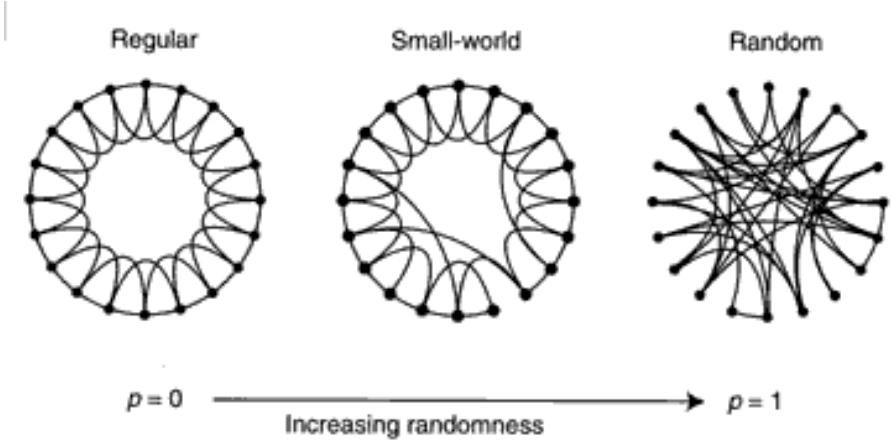
The Watts and Strogatz Model

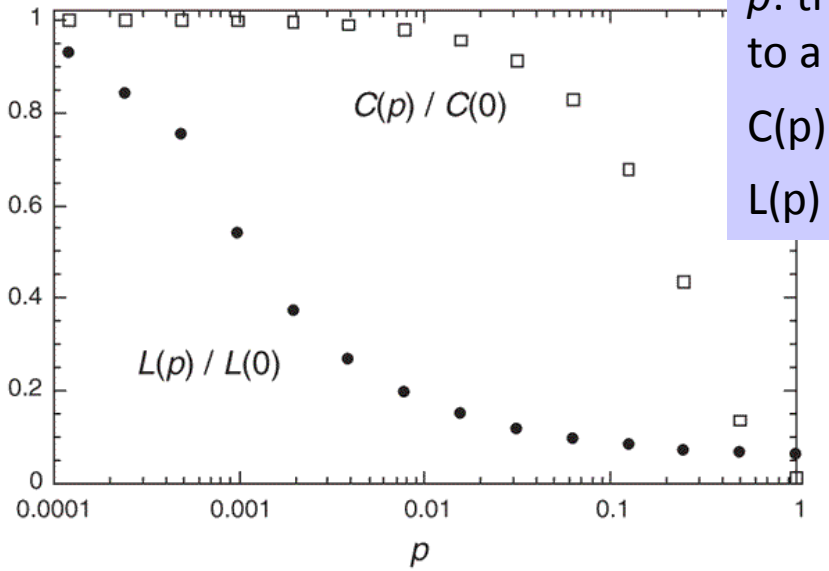
- A random network has low local clustering coefficient:
  - $$C_i = \frac{\langle L_i \rangle}{k_i \times (k_i - 1)/2} = p = \frac{\langle k_i \rangle}{N}$$

L_i: Expected # of links between the node i's k_i neighbors

Possible # of links between the node i's k_i neighbors

- The Watts and Strogatz Model (aka: small world model) Duncan J. Watts & Steven Strogatz, Nature 1998)
  - Interpolates between regular lattice and a random network to generate graphs with
    - **Small-world**: short average path lengths
    - **High clustering coefficient:**





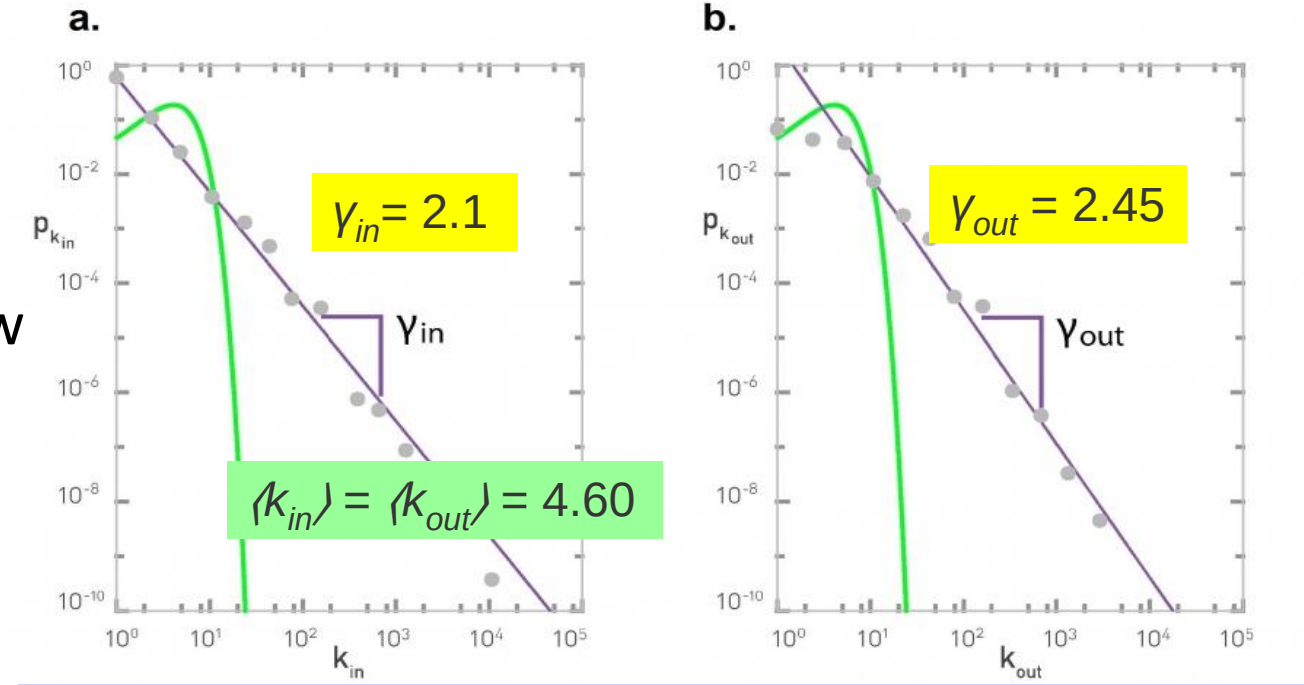
p: the prob. each link is rewired to a randomly chosen node

C(p) : clustering coeff.

L(p) : average path length

Real Networks Are Not Random

- The random model does not fit many real networks, e.g., WWW
- Mapping WWW into log-log plot: It follows a power-law not Poisson distribution
- A scale-free network is a network whose degree distribution follows a power law
- Power law vs. Poisson distributions:
  - For small k, power law is above Poisson: more small-degree nodes
  - For k in the vicinity of <k>, power law is below Poisson: an excess of nodes with k ≈ <k> in a random network
  - For large k, power law is above Poisson: prob. of observing hubs is orders of magnitude higher in a scale-free than in a random network



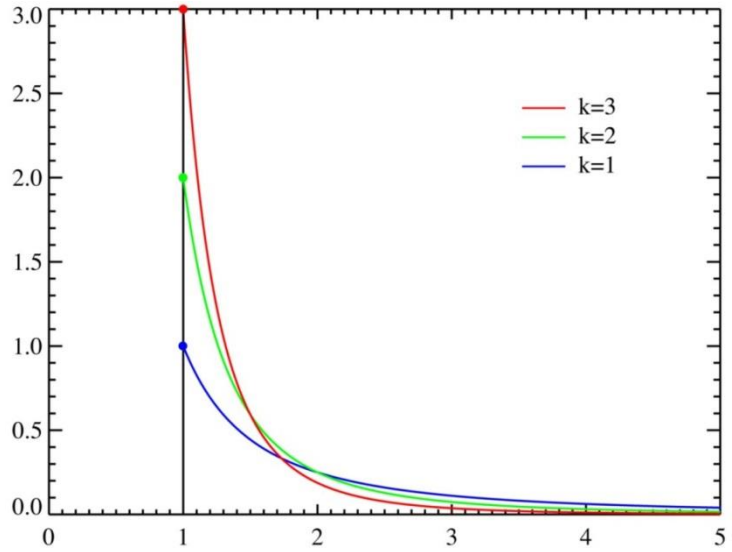
Degree distribution of WWW *vs.* that predicted by Poisson (green)
R. Albert, H. Jeong, A-L Barabasi, Nature, **401** 130 (1999)

Common Properties of Real Networks

- Real network: **large, sparse** (# of edges |E| = O(n), n: # of nodes)
- **Small-world property**: Avg. path length μ_L scales logarithmically with n (# of nodes in the graph): $\mu_L \propto \log n$
  - Ultra-small-world property: $\mu_L \ll \log n$
- **Scale-free property** (**power law distribution**): most nodes have very small degree, but a few hub nodes have high degrees
  - The probability that a node has degree k: $f(k) \propto k^{-\gamma}$
  - log-log plot shows a straight line: $\log f(k) = \log(\alpha k^{-\gamma}) = -\gamma \log k + \log \alpha$
- **Clustering effect**:
  - Two nodes are more likely to be connected if they share a common neighbor
  - Clustering effect: a high clustering coefficient for graph G
  - C(k): avg clustering coefficient for nodes with degree k
  - Power law relationship between C(k) and k: $C(k) \propto k^{-\gamma}$

Power Law (or Pareto) Distributions

- Pareto distribution (heavy-tailed, or power law):
  - Pareto(x|k, m) = k m^k $x^{-(k+1)}$ **I**($x \geq m$)
  - x must be greater than some constant m but not too much greater
  - Distributions: mode = m
    - mean = km/(k-1) if k > 1
    - variance = $m^2k/((k-1)^2(k-2))$ if k > 2



- For variables assuming integer values > 0, probability of value $x \sim 1/x^{\alpha}$
  - This is why it is called *power law* distribution, also referred to as *scale-free*
  - If we plot the distribution on a log-log scale, it forms a straight line
  - Typically $0 < \alpha < 2$; smaller α gives heavier tail
- Why long tails or heavy tails? For binomial, normal, and Poisson distributions, the tail probabilities approach 0 *exponentially* fast
- What kind of phenomena does this distribution model?
  - Word frequency vs their rank (the, of, …); wealth distribution; …

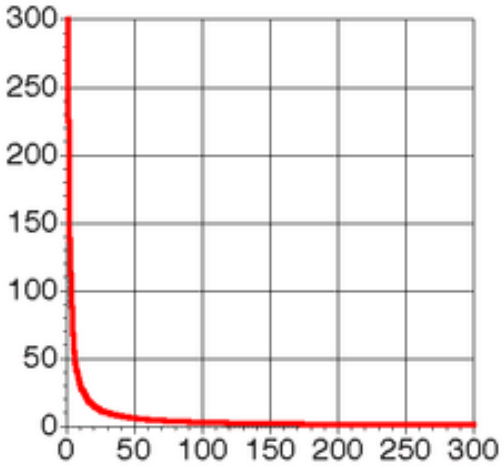
Distinguishing Distributions in Data

- 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
  - Get *count* of observed data in equal-sized bins: look at counts on a *log-log plot*
- Power law: $\log(P(X = x)) = \log(1/x^{\alpha}) = -\alpha \log(x)$
  - linear, slope $-\alpha$
- Normal: $\log(P(X = x)) = \log(\alpha \exp(-x^2/b)) = \log(\alpha) - x^2/b$
  - non-linear, concave near mean
- Poisson: $\log(P(X = x)) = \log(\exp(-\lambda)\, \lambda^x/x!)$
  - also non-linear

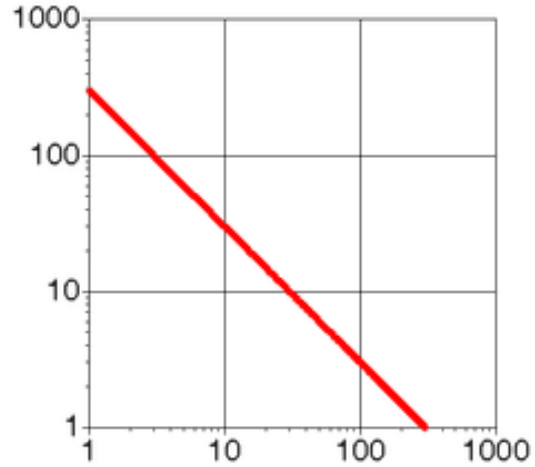
Zipf's Law and Zipf Distribution

- Pareto distribution vs. Zipf's Law
  - Pareto distributions are continuous probability distributions
  - Zipf's law: a discrete counterpart of the Pareto distribution
- Zipf's law example:
  - Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table
  - The most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc.
- Example: *rank* events by their *frequency of occurrences.* The resulting distribution often is a power law!
- Other examples: North America city sizes, personal income, file sizes, genus sizes (number of species)

A Zipf distribution with 300 data points



Linear scales on both axes



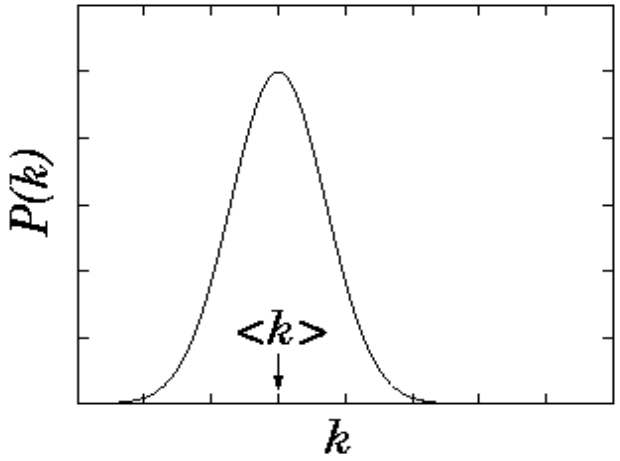
Log scales on both axes

Discovered by Examining the Real World…

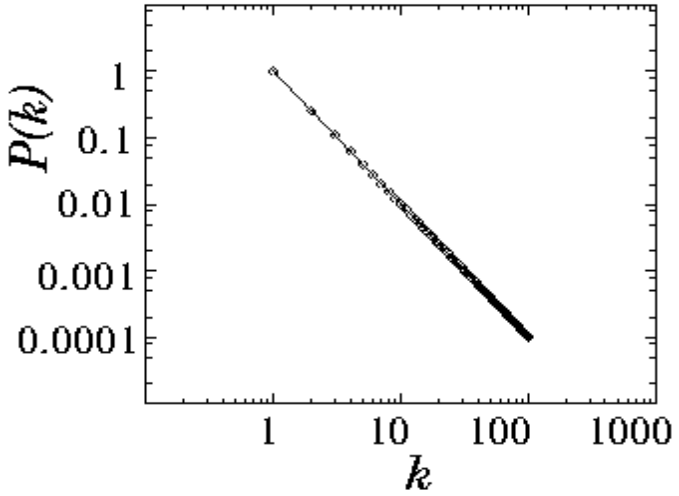
- Watts examines three real networks as case studies:
  - the Kevin Bacon graph
  - the Western states power grid
  - the C. elegans nervous system
- For each of these networks, he:
  - computes its size, diameter, and clustering coefficient
  - compares diameter and clustering to *best* Erdos-Renyi approx.
  - shows that the *best* α-model approximation is better
  - important to be "fair" to each model by finding best fit
- Overall,
  - if we care only about diameter and clustering, α is better than p

Not All the Networks Are Scale-Free

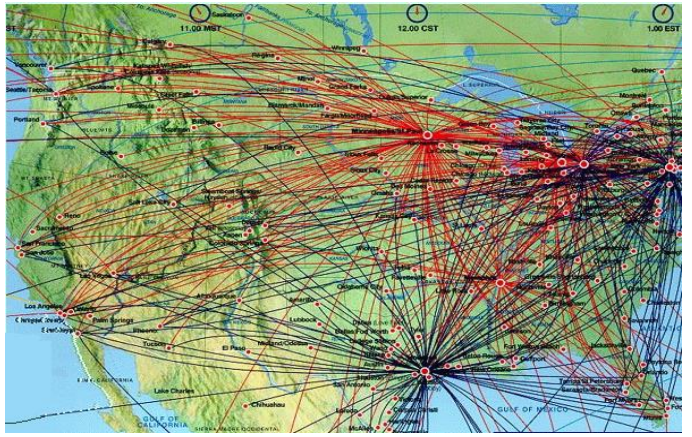
Poisson distribution



Power-law distribution



Highway Network



Scale-free Network

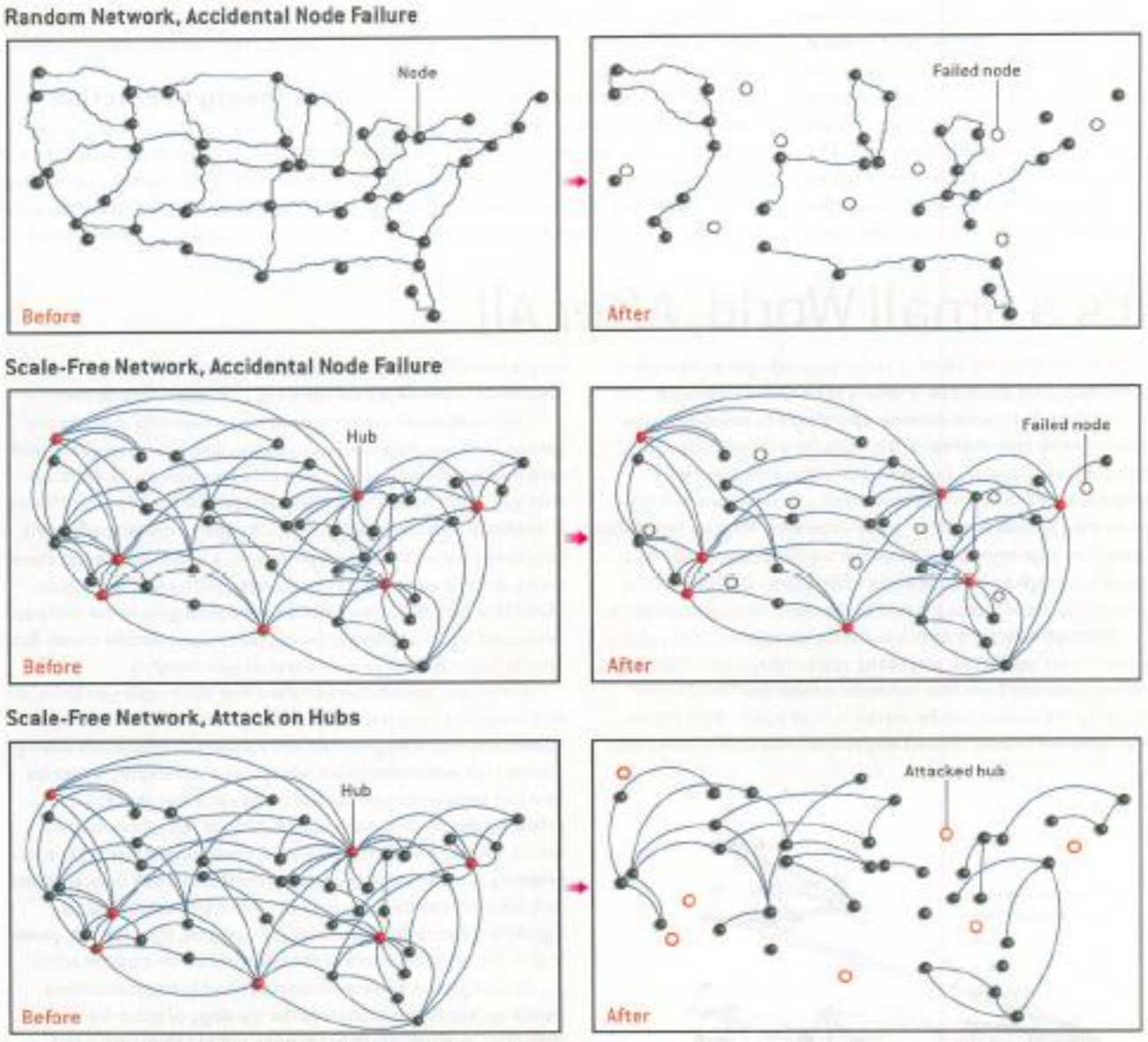
Barabási-Albert Model: Preferential Attachment

- Major limitation of the Watts-Strogatz model
  - It produces graphs that are homogeneous in degree
  - Real networks are often inhomogeneous in degree, having hubs and a scale-free degree distribution (*scale-free networks*)
- **Scale-free networks** are better described by the ***preferential attachment*** family of models, e.g., the *Barabási–Albert (BA) model*
  - *"rich-get-richer":* New edges are more likely to link to nodes with higher degrees
  - **Preferential attachment:** The probability of connecting to a node is proportional to the current degree of that node
- This leads to the proposal of a new model: ***scale-free network***, a network whose degree distribution follows a ***power law***, at least asymptotically

Scale-Free Networks: Major Ideas

- 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

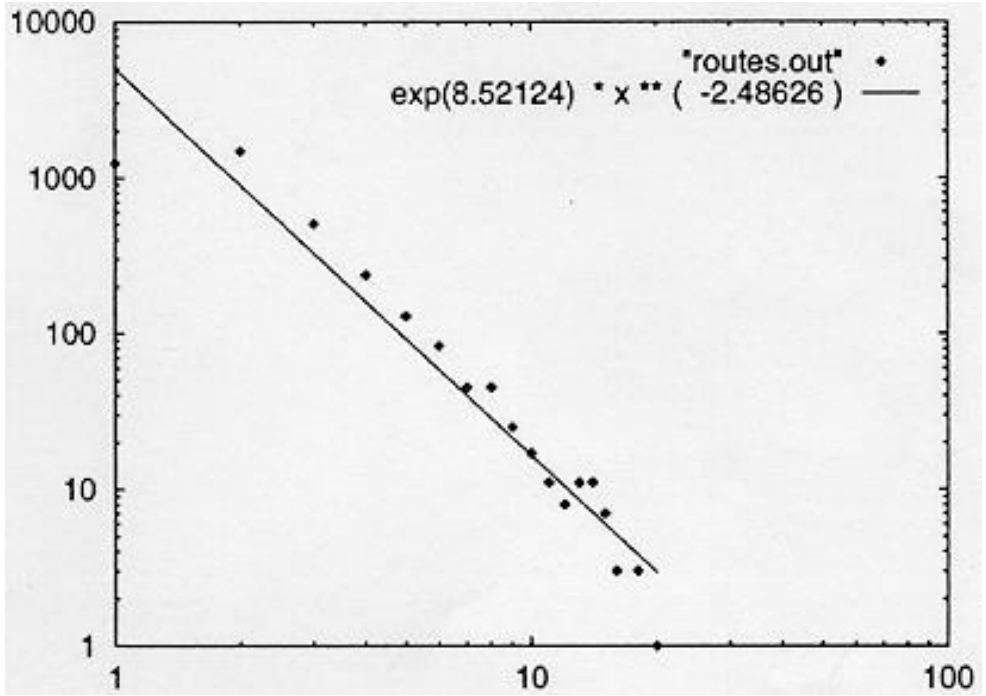
Robustness of Random vs. Scale-Free Networks



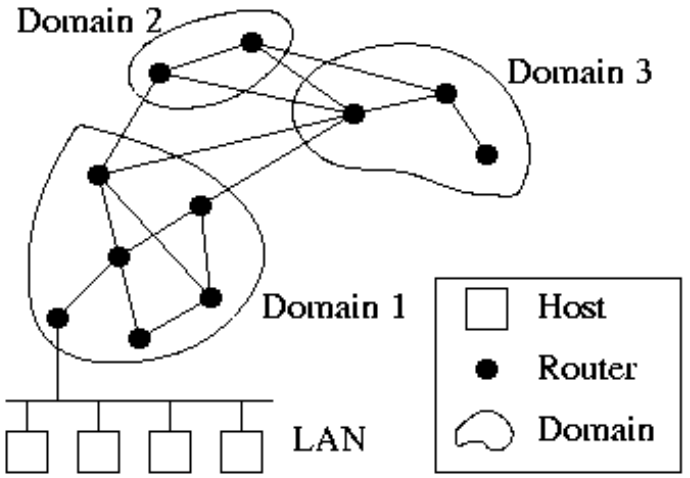
- The accidental failure of a number of nodes in a random network can fracture the system into non-communicating islands
- Scale-free networks are more robust in the face of such failures
- Scale-free networks are highly vulnerable to a coordinated attack against their hubs

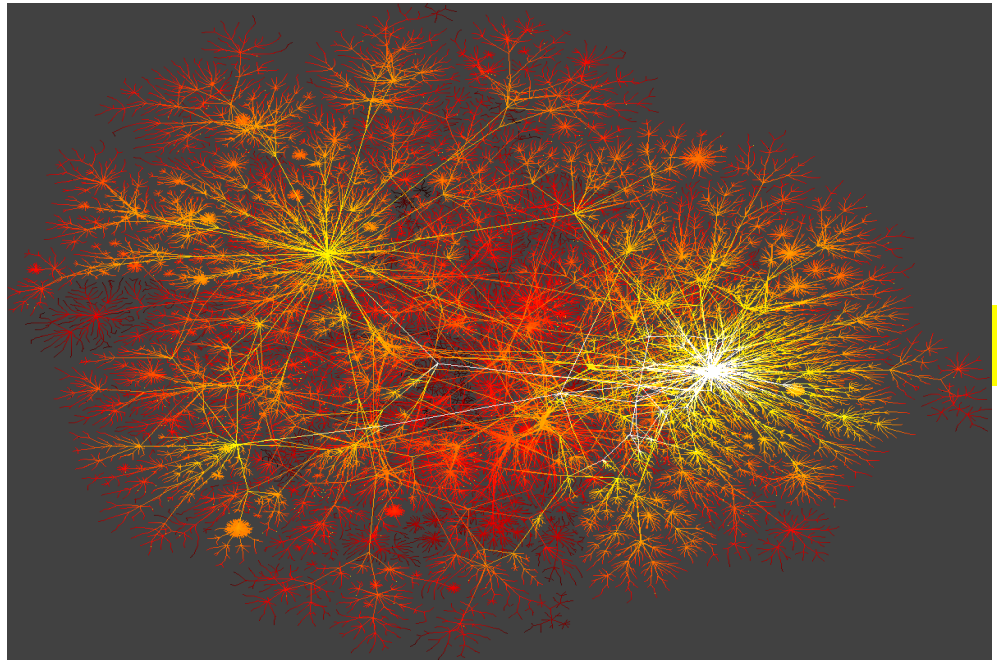
Real World Case 1: Internet Backbone

Nodes: computers, routers
Links: physical lines



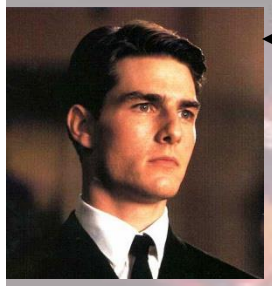
(Faloutsos, Faloutsos and Faloutsos, 1999)



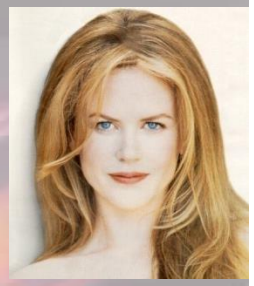


Internet-Map

Real World Case: Actor Connectivity



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



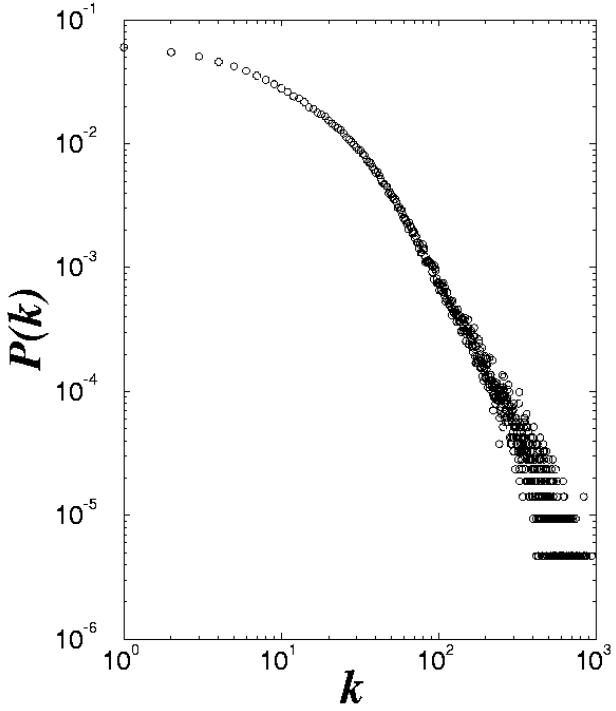
N = 212,250 actors

⟨k⟩ = 28.78

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

γ=2.3

Nodes: actors
Links: cast jointly



Real World Case: Science Citation Index

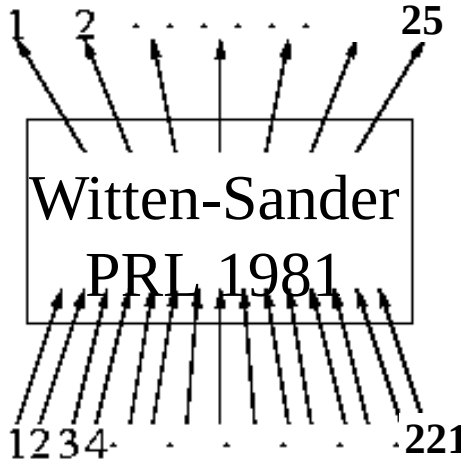
1,000 Most Cited Physicists, 1981-June 19

Out of over 500,000 Examined

(see http://www.sst.nrel.gov)

Author name		Institute	Country	Field	avg. cites			
Witten	E	Princeton (U)	USA, NJ	High-energy (T)	168			
Gossard	AC	UCSB (U)	USA, CA	Semiconductors (E)	41			
Cava	RJ	Bell Labs (I)	USA, NJ	Superconductors (E)	65			
Batlogg	B	Bell Labs (I)	USA, NJ	Superconductors (E)	83			
Ploog	K	Max-Planck (NL)	Germany	Semiconductors (E)	19			
Ellis	J	Euro Nuclear Cent.	Switzerland	Astrophysics (E)	40			
Fisk	Z	Florida State (U)	USA, FL	Solid State (E)	23			
Cardona	M	Max Planck (NL)	Germany	Semiconductors (E)	20			
Nanopoulos	DV	Texas A&M (U)	USA, TX	High-energy (E)	39			
Heeger	AJ	UCSB (U)	USA, CA	Polymers (E)	34			
Lee*	PA				73			
Suzuki*	T				7.6			
Anderson	PW	Princeton (U)	USA, NJ	Solid State (T)	80			
Suzuki*	M				12			
Freeman	AJ	Northwestern (U)	USA, IL	Solid State (T)	27	389	10411	15
Tanaka*	S				11	963	10404	16
Muller	KA	Zurich Univ. (U)	Switzerland	Superconductivity (E)	82	122	10049	17
Schneemeyer	LF	Bell Labs (I)	USA, NJ	Superconductivity (E)	63	156	[illegible]	18
Chemla	DS	USB (U/NL)	USA, CA	Optics (E)	60	162	[illegible]	[illegible]
Morkoc	H	U. of IL (U)	USA, IL	Semiconductors (E)	20	477	9652	19
Miller	DAB	Stanford (U)	USA, CA	Semiconductors (E)	67	144	9652	21
Chu	CW	Houston Univ. (U)	USA, TX	Superconductivity (E)	44	213	9453	22
Bednorz	JG	IBM (I)	Switzerland	Superconductivity (E)	110	85	9311	23
Cohen	ML	UCB/LBL (U/NL)	USA, CA	Solid State (T)	33	284	9311	23
Meng	RL	Houston Univ. (U)	USA, TX	Superconductivity (E)	86	108	9306	[illegible]
Waszczak	JV	AT&T (I)	USA, NJ	Superconductivity (E)	57	162	9176	26
Shirane	G	Brookhaven (NL)	USA, NY	Superconductivity (E)	33	269	8841	27
Wiegmann	W	Bell Labs (I)	USA, NJ	Semiconductors (E)	85	104	8822	28
Vandover	RB	Bell Labs (I)	USA, NJ	Magnetism (E)	67	129	8686	29
Uchida*	S				28	301	8520	30
Hor	PH	Houston Univ. (U)	USA, TX	Superconductivity (E)	72	119	8512	31
Murphy	DW			Astronomy (E)	111	76	8439	32
Birgeneau	RJ	MIT (U)	USA, MA	Superconductivity (E)	41	286	8375	33
Jorgensen	JD	Argonne (NL)	USA, IL	Superconductivity (E)	50	167	8298	34
Hinks	DG	Argonne (NL)	USA, IL	Superconductivity (E)	37	223	8263	35

* citation total may be skewed because of multiple authors with the same name



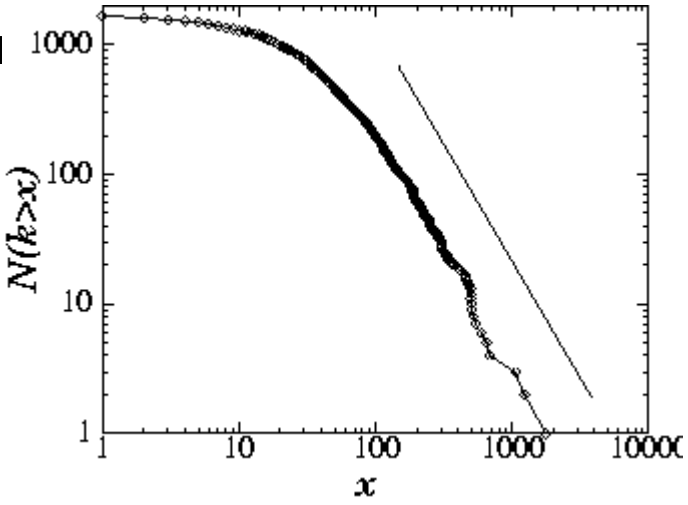
Nodes: papers

Links: citations

1736 PRL papers (1988)

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

$(\gamma = 3)$

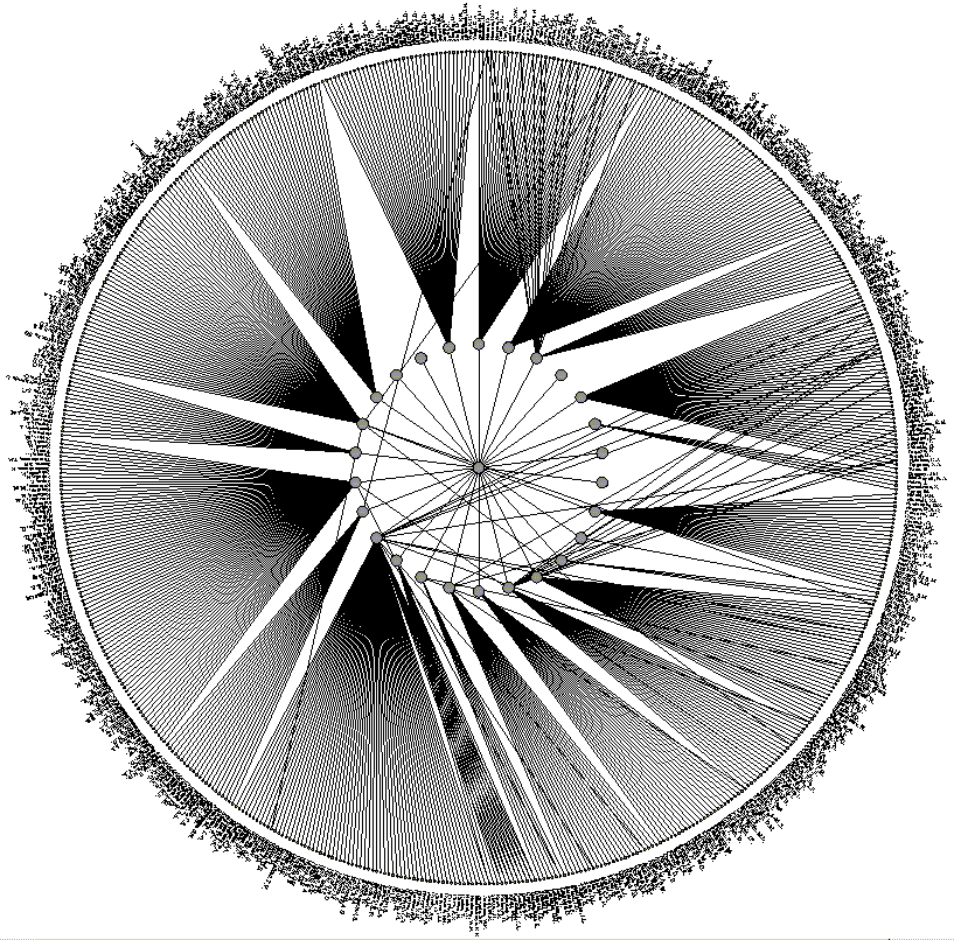


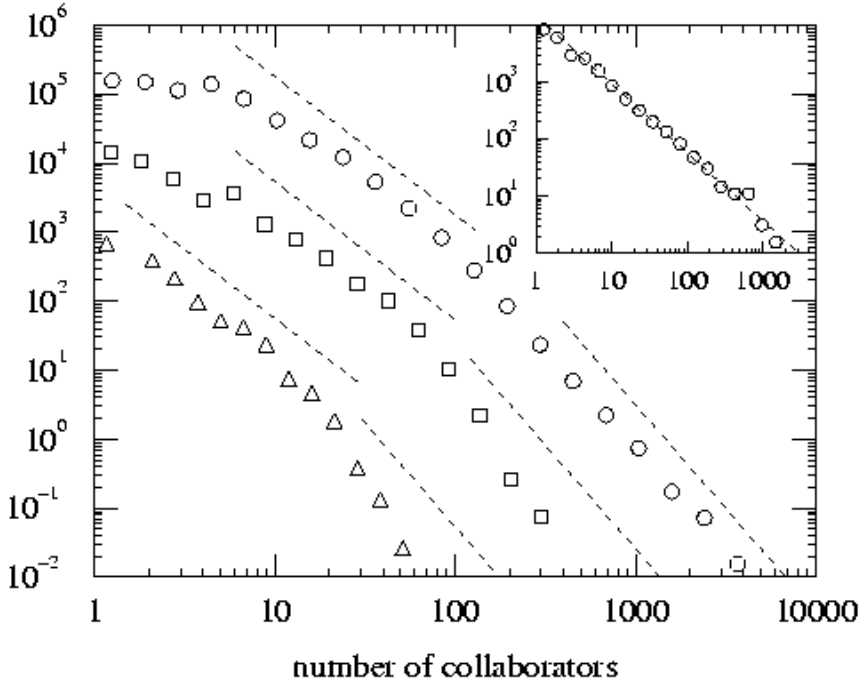
(S. Redner, 1998)

Real World Case: Science Co-authorship

Nodes: scientist (authors)

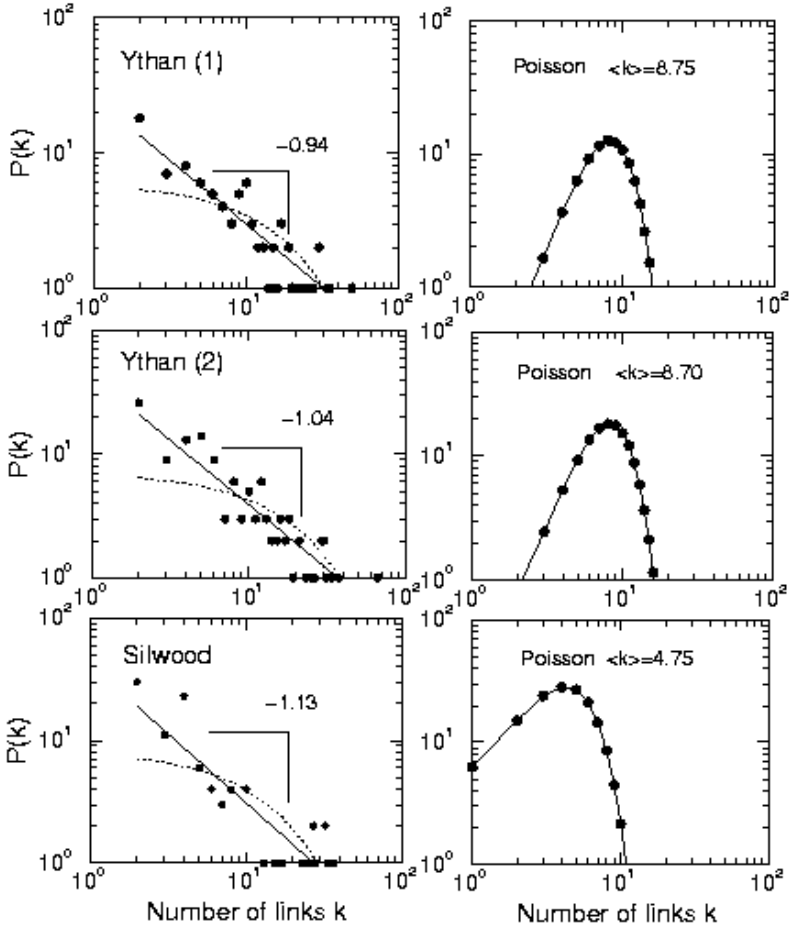
Links: write paper together





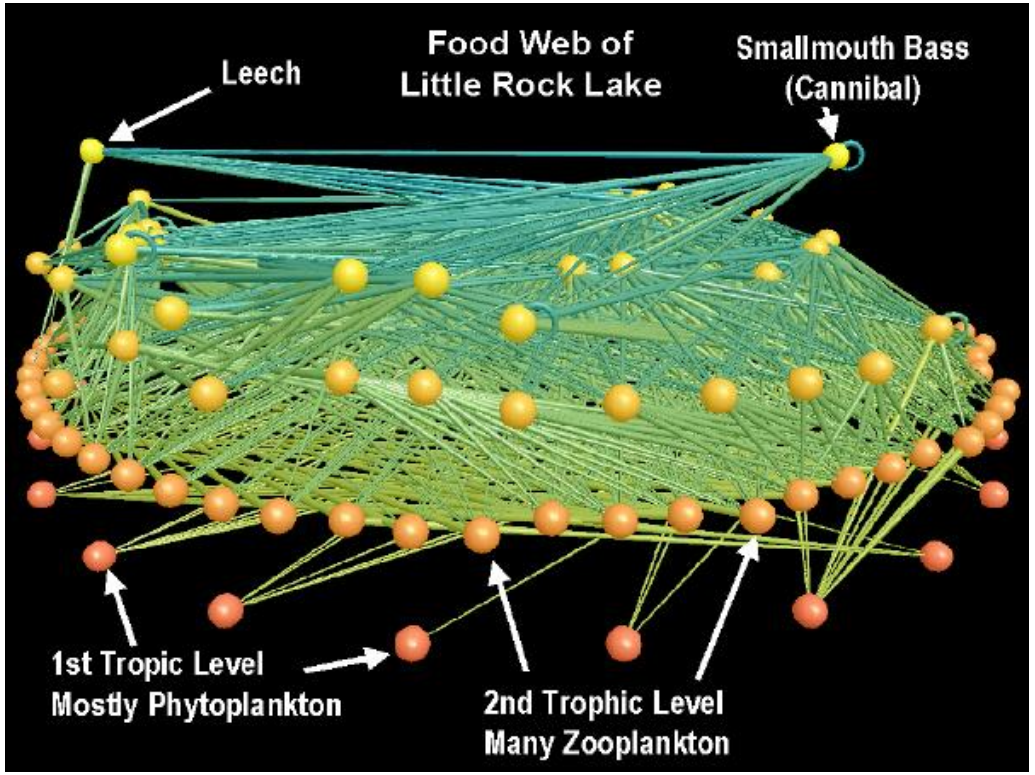
(Newman, 2000, H. Jeong et al 2001)

Real World Case: Food Web



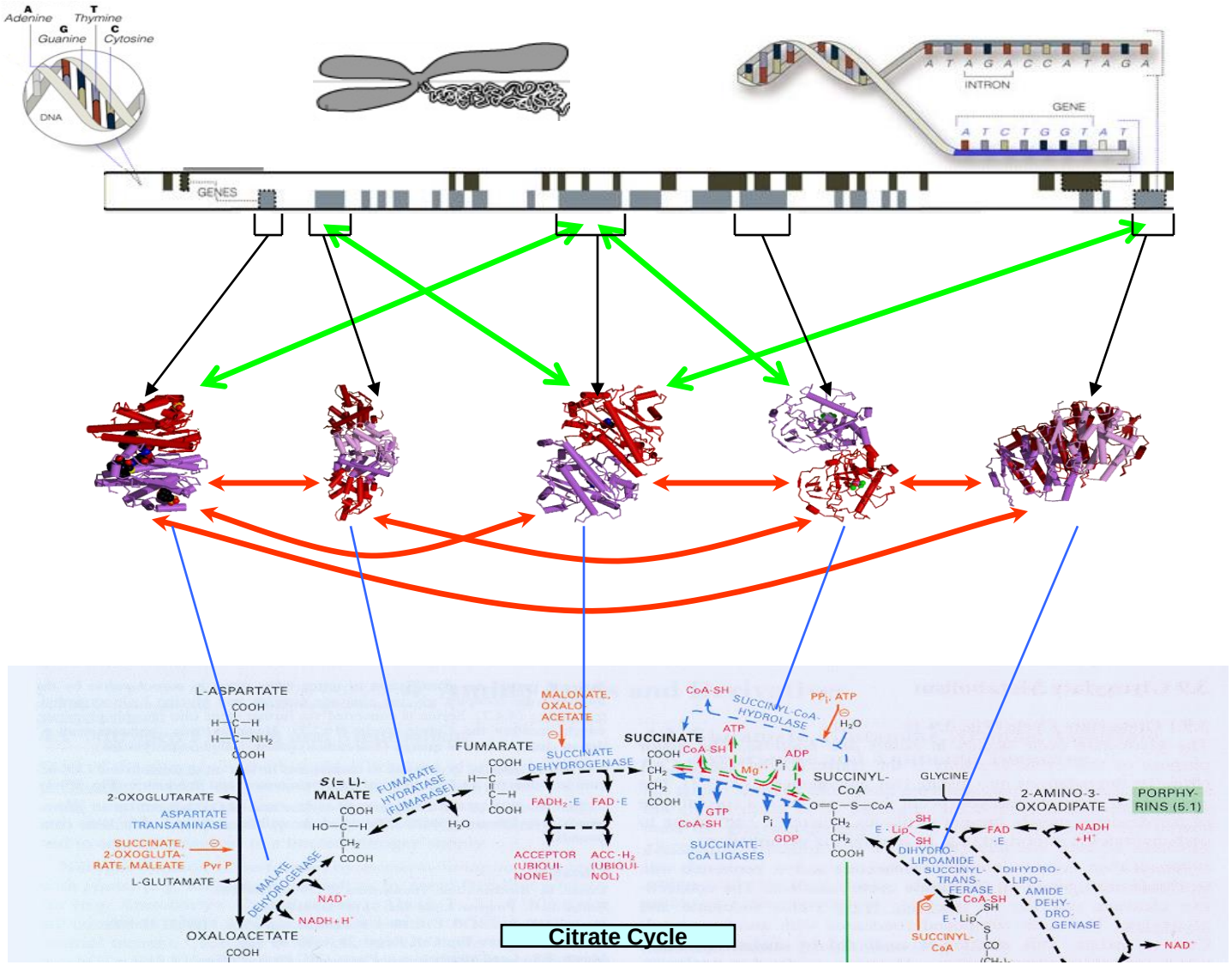
R. Sole (cond-mat/0011195)

Nodes: trophic species
Links: trophic interactions



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

Real World Case: Biology Map



GENOME

protein-gene interactions

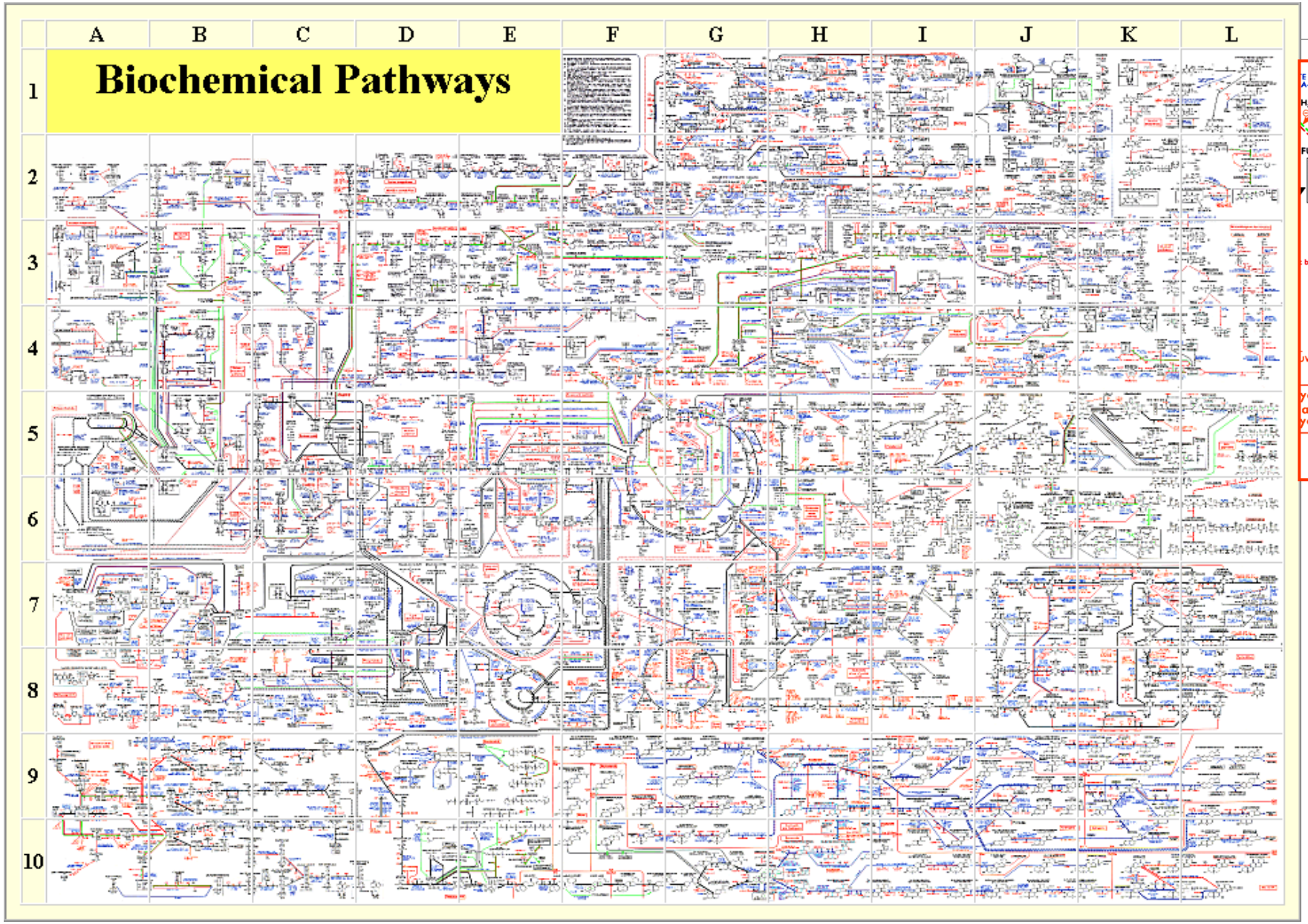
PROTEOME

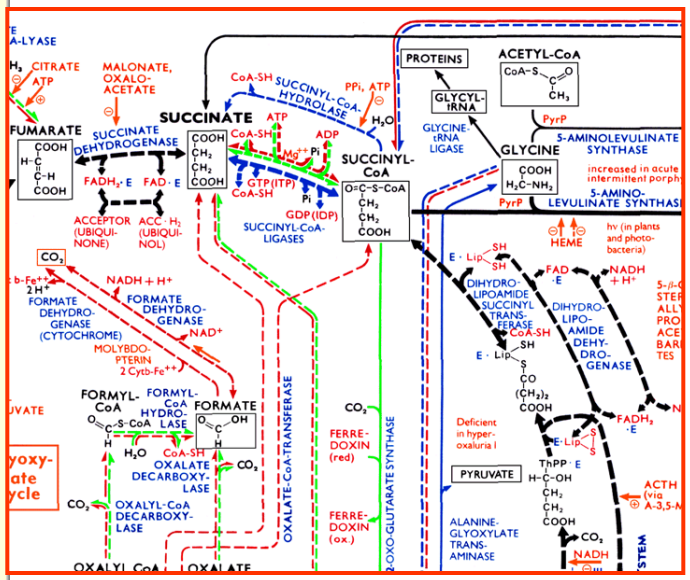
protein-protein interactions

METABOLISM

Bio-chemical reactions

Biological Pathways and Complex Biological Networks

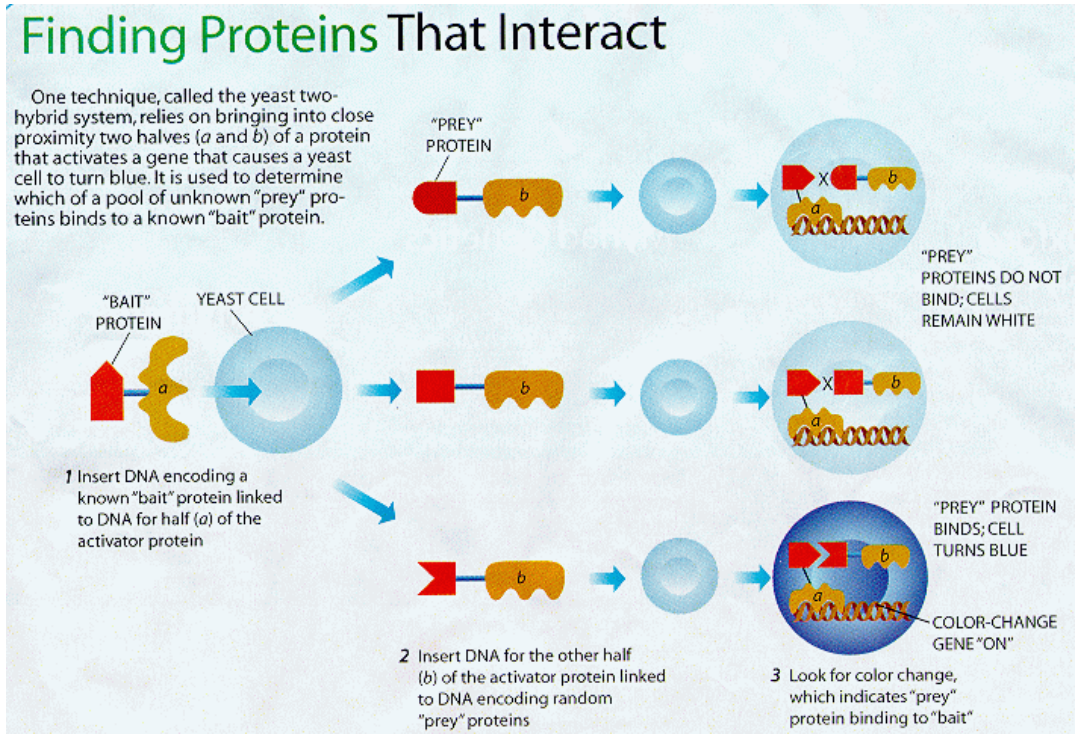


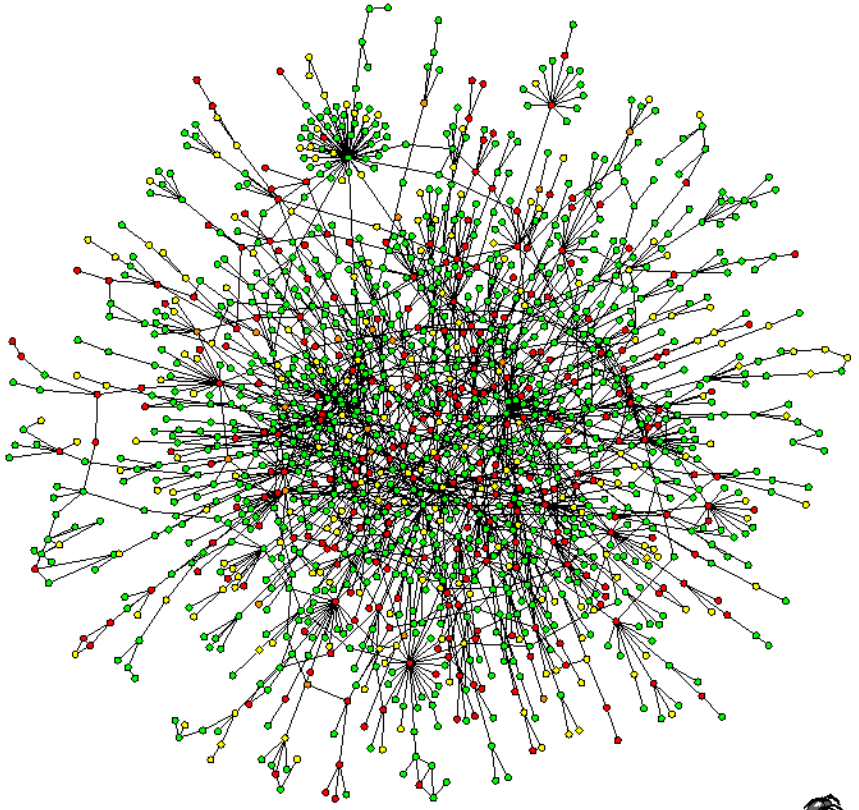


Protein Interaction Map: Yeast Protein Network

Nodes: proteins

Links: physical interactions (binding)





P. Uetz, *et al.*, *Nature* **403**, 623-7 (2000)

Patterns in Static vs. Dynamic Networks

- Static graphs
  - Skewed/power-law degree distribution
  - Power-law eigenvalue distribution
  - Small diameter
  - Triangle power law
  - Triangle-degree power law
- Dynamic graphs
  - Shrinking diameter
  - Densification power law
  - Jelling point
  - Oscillating size of NLCC (non-largest connected components)
  - Principal eigenvalue over time

Based on D. Chakrabarti and C. Faloutsos, Graph Mining: Laws, Tools, and Case Studies, Morgan & Claypool, 2012

Patterns in Weighted Networks

- Weighted Static Graphs
  - Superlinear weight-degree relationship ('snapshot power law')

- Weighted Dynamic Graphs
  - WPL –Weight Power Law. Super-linear relationships, over time between total graph weight *W(t)*, total edge count *E(t)* etc.
  - WLPL – principal eigenvalue of weighted graph follows a power law over time, with respect to edge count

Based on D. Chakrabarti and C. Faloutsos, Graph Mining: Laws, Tools, and Case Studies, Morgan & Claypool, 2012

Modeling Network Evolution

- Densification power law
  - The # of edges grows super-linearly (i.e., more than linearly) to # of nodes, following a power law, with a positive *densification exponent*

$$E(t) \propto N(t)^{\beta}$$ where β = 1.03~1.07 in many real graphs

- Shrinking diameter: The diameter of the graph *shrinks* as a graph grows over time
- The Forest-Fire model: A preferential-attachment model that matches the densification power law and the shrinking diameter patterns of graph evolution
  - The graph grows one node at a time. The new node *v* adds links to the existing node according to a "forest fire" process
    - Pick an ambassador node *w* uniformly at random and the links to *w*
    - Select some of ambassador's edges, and follow these edges and repeat
      - Similar to capture a "forest fire" at *w* and spread to other nodes

- J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graphs over time: Densification laws, shrinking diameters and possible explanations. KDD'05

Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks
- Modeling of Network Formation
- Primitives of Social Networks 
- 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

Web 2.0 Examples

- Blogs
  - Blogspot
  - Wordpress
- Wikis
  - Wikipedia
- Social Networking Sites
  - Facebook
  - Orkut
- Digital media sharing websites
  - Youtube
  - Flickr
- Social Tagging
  - Del.icio.us
- Others
  - Twitter
  - Yelp

Friendship Networks vs. Blogosphere

Friendship Networks	Blogosphere
Explicit Links/Edges	Implicit Links/Edges
Undirected Graph	Directed Graph
Network Centrality Measures	Blog Statistics
Quantifying Spread of Influence	Quantifying Influential Members
Nodes are members/actors	Nodes can be bloggers/blogs or blog sites
Strictly defined graph structure	Loosely defined graph structure
"Being in touch" or "Making Friends"	Sharing ideas and opinions
Person-to-person	Person-to-group
Friendship Oriented	Community Oriented
Member's Reputation/Trust based on network connections and/or location in the network	Member's Reputation/Trust based on the response to other member's knowledge solicitations

Adapted from H. Liu & N. Agarwal, KDD'08 tutorial

Society: Social Networks

Nodes: individuals

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

Six Degrees of Separation

S. Milgram (1967)

John Guare

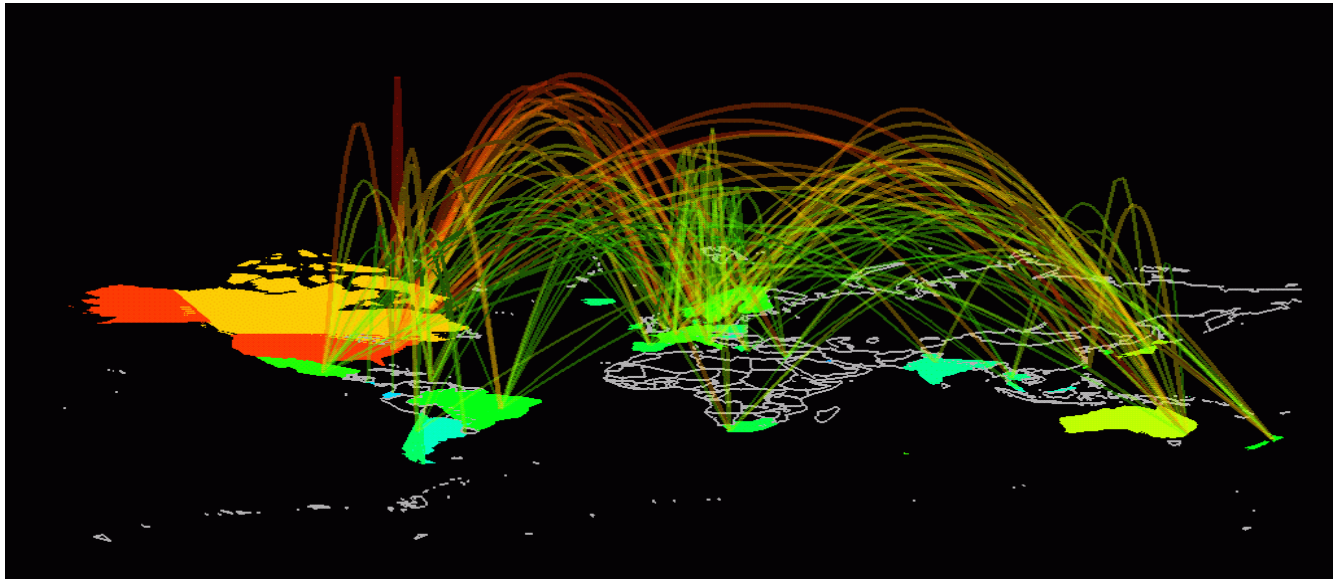
Social networks: Many individuals with diverse social interactions between them

Communication Networks

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

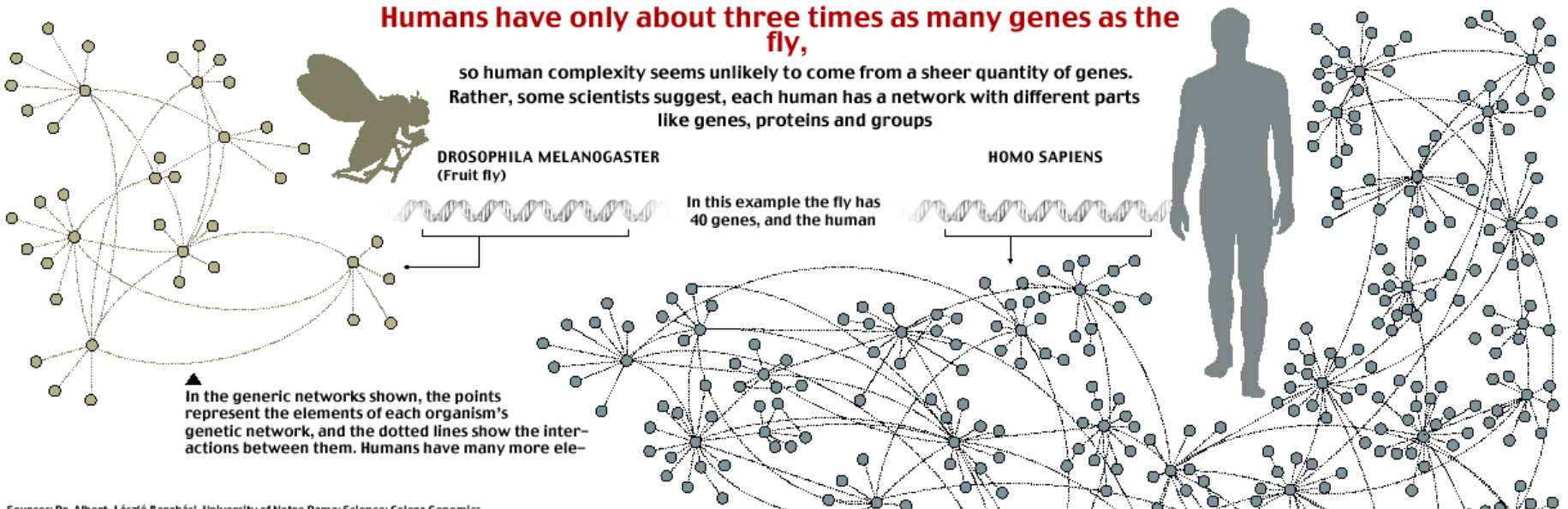
- Computers
- Routers
- Satellites

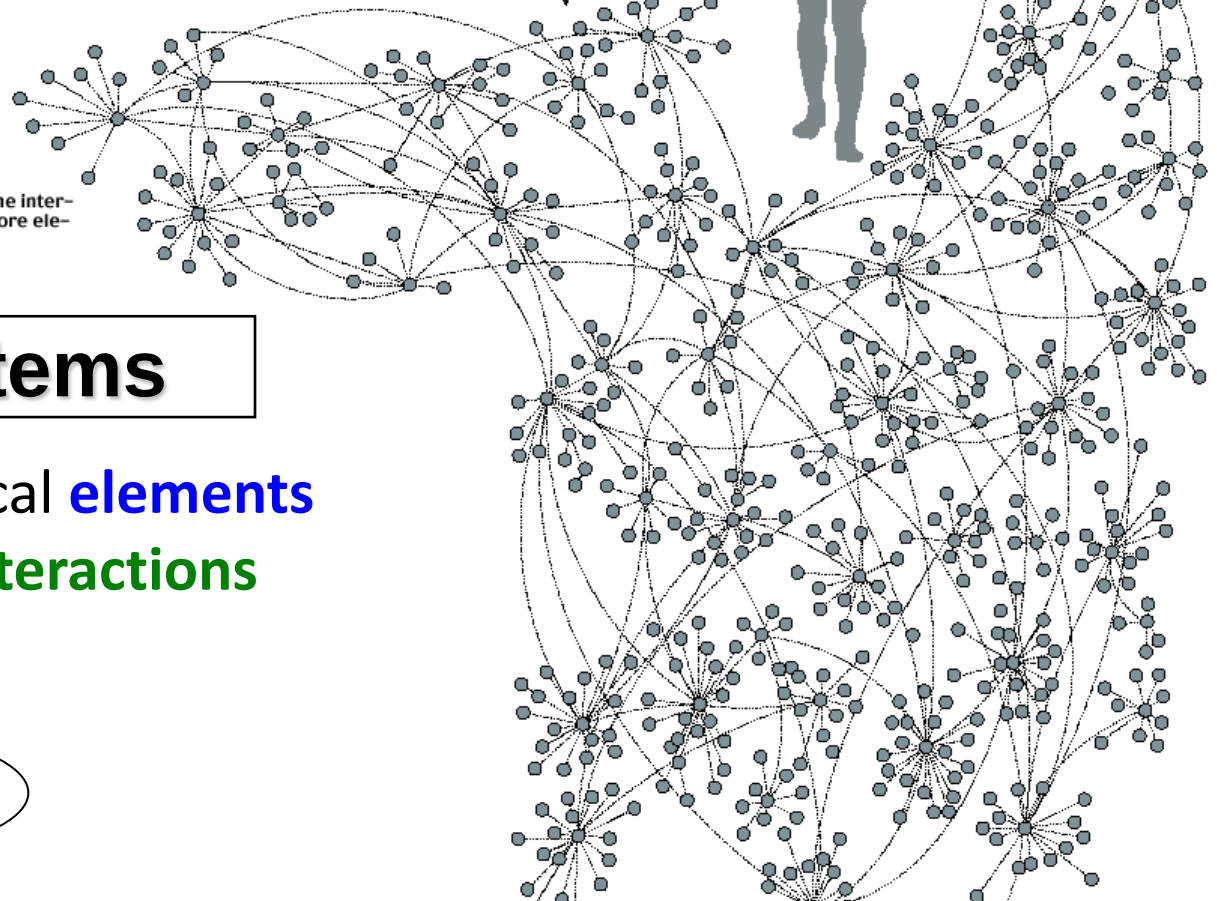
- Phone lines
- TV cables
- EM waves



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

Biological Networks





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, organic growth: vertices "decide" who to link to
  - interaction restricted to links
  - mixture of local and long-distance connections
  - abstract notions of distance: geographical, content, social,…
- *Social network theory and link analysis*
  - Do natural networks share more *quantitative* universals?
  - What would these "universals" be?
  - How can we make them precise and measure them?
  - How can we explain their universality?

Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks
- Modeling of Network Formation
- Primitives of Social Networks
- Summary 

Some Typical Network Data Sets

- Collaboration graphs
  - Co-authorships among authors
  - co-appearance in movies by actors/actresses
- Who-Talks-to-Whom graphs
  - Microsoft IM (Instant-Messaging)-graphs
- Information Linkage graphs
  - Web, citation graphs
- Technological graphs
  - Interconnections among computers
  - Physical, economic networks
- Networks in the Natural World
  - Food Web: who eats whom
  - Neural connections within an organism's brain
  - Cells metabolism

Summary

- Primitives for networks
- Measure and metrics of networks
  - Degree, eigenvalue, Katz, PageRank, HITS
- Models of network formation
  - Erdös-Rényi, Watts and Strogatz, scale-free
- Social networks

References: Introduction to Networks

- S. Brin and L. Page, The anatomy of a large scale hypertextual Web search engine. WWW7.
- S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, Mining the link structure of the World Wide Web. IEEE Computer'99
- D. Cai, X. He, J. Wen, and W. Ma, Block-level Link Analysis. SIGIR'2004
- P. Domingos, Mining Social Networks for Viral Marketing. IEEE Intelligent Systems, 20(1), 80-82, 2005
- D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Cambridge Univ. Press, 2010
- L. Getoor: Lecture notes from Lise Getoor's website: www.cs.umd.edu/~getoor/
- D. Kempe, J. Kleinberg, and E. Tardos, Maximizing the Spread of Influence through a Social Network. KDD'03
- J. M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, J. ACM, 1999
- D. Liben-Nowell and J. Kleinberg. The Link Prediction Problem for Social Networks. CIKM'03
- M. Newman, Networks: An Introduction, Oxford Univ. Press, 2010
- D. Chakrabarti and C. Faloutsos, Graph Mining: Laws, Tools, and Case Studies, Morgan & Claypool, 2012

Probability and Random Variables (A Brief Review)

- A *random variable* X is simply a variable that *probabilistically* assumes values in some set
  - set of possible values sometimes called the *sample space* S of X
  - Sample space may be small and simple or large and complex
    - S = {Heads, Tails}, X is outcome of a coin flip
    - S = {0,1,…,U.S. population size}, X is number voting democratic
    - S = all networks of size N, X is generated by *preferential attachment*
- Behavior of X determined by its *distribution* (or *density*)
  - For each value x in S, specify p(X = x): these probabilities sum to exactly 1 (mutually exclusive outcomes)
  - complex sample spaces (such as large networks):
    - distribution often defined *implicitly* by simpler components
    - might specify the probability that each *edge* appears independently
    - this *induces* a probability distribution over *networks*
    - may be difficult to *compute* induced distribution

Independence, Expectation & Variance

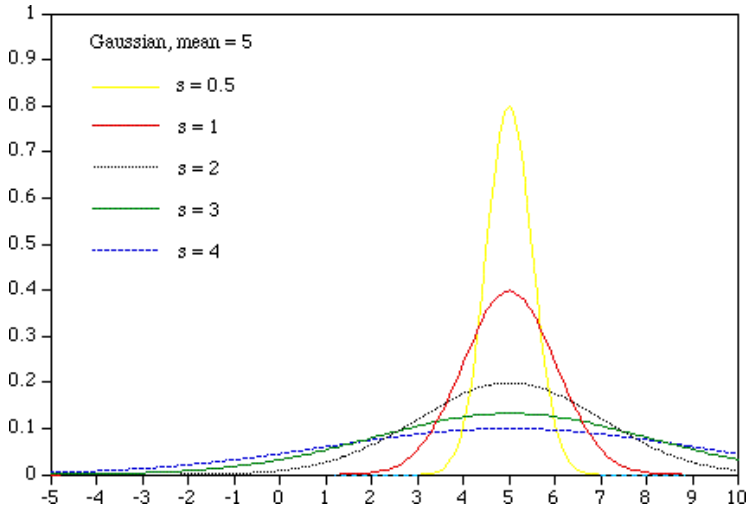
- *Independence*:
  - let X and Y be random variables.
  - unconditional independence: for any x & y, $p(X = x, Y = y) = p(X=x) \times p(Y=y)$
  - intuition: value of X does not influence value of Y, vice-versa
  - conditional independence: $p(X, Y \mid Z) = p(X \mid Z)\ p(Y \mid Z)$
- *Expected (mean) value* of X: μ
  - only makes sense for *numeric* random variables
  - "average" value of X according to its distribution
  - formally, $E[X] = \sum_{x \in X} x\ p(x)$, i.e., sum over all x in X
  - *always* true: $E[X + Y] = E[X] + E[Y]$
  - true only for *independent* random variables: $E[XY] = E[X]E[Y]$
- *Variance* of X:
  - $Var[X] = E[(X - \mu)^2]$; often denoted by σ^2
  - *standard deviation* $\sigma = sqrt(Var[X])$
- *Union bound:*
  - for any X, Y, $p(X=x, Y=y) \leq p(X=x) + p(Y=y)$

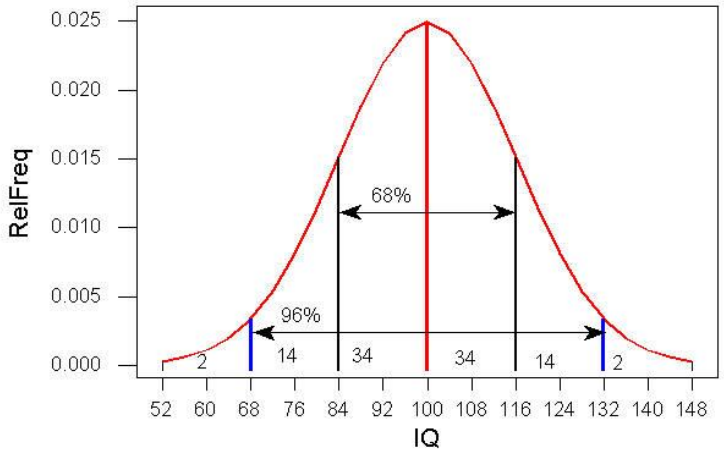
Convergence to Expectations

- Let $X_1, X_2, \ldots, X_n$ be:
  - *independent* random variables
  - with the *same* distribution p(X=x)
  - expectation $\mu = E[X]$ and variance σ^2
  - independent and identically distributed (i.i.d.)
  - essentially n repeated "trials" of the same experiment
  - natural to examine random variable $Z = (1/n) \sum_{i=1:n} X_i$
  - example: number of heads in a sequence of coin flips
  - example: degree of a vertex in the random graph model
  - E[Z] = E[X]; what can we say about the *distribution* of Z?
- *Central Limit Theorem:*
  - as n becomes large, Z becomes *normally distributed*
    - with expectation μ and variance σ^2/n

The Gaussia (Normal) Distribution

- The *normal* or *Gaussian* density applies to continuous, real-valued random variables
- characterized by mean μ and std. deviation σ
- *Density* at x is defined as $(1/(\sigma \sqrt{(2\pi)})) \exp(-(x-\mu)^2/2\sigma^2)$
  - Special case μ = 0, σ = 1: $\alpha \exp(-x^2/\beta)$ for some constants α, β > 0
  - peaks at x = μ, then dies off *exponentially* rapidly
- The classic "bell-shaped curve", e.g., exam scores, human body temperature
- Remarks:
  - can control mean and standard deviation independently
  - can make as "broad" as we like, but always have *finite variance*





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