Bioinformatics: Network Analysis

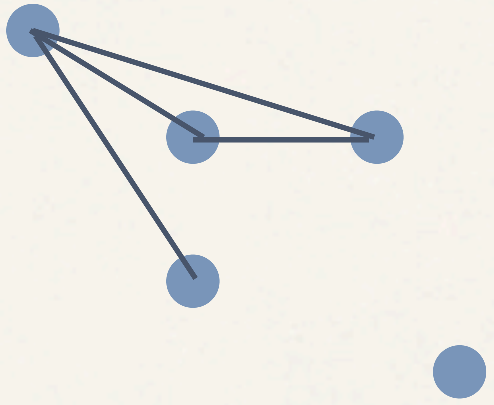
Graph-theoretic Properties of Networks

COMP 572 (BIOS 572 / BIOE 564) - Fall 2013

Luay Nakhleh, Rice University

Graphs

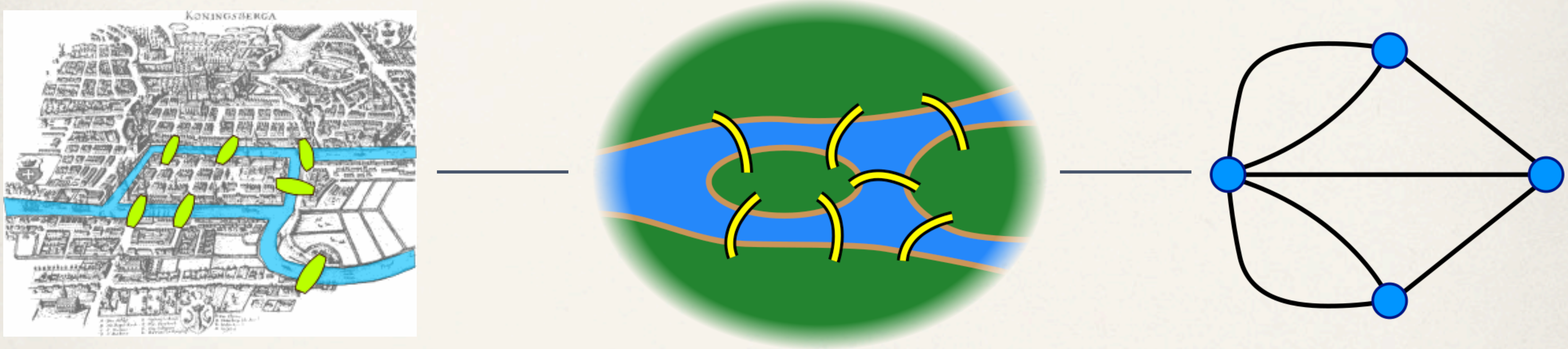
- A graph is a set of *vertices*, or *nodes*, and *edges* that connect pairs of vertices



Example: a graph with 5 vertices and 4 edges

The Königsberg Bridge Problem

Is it possible to walk a route that crosses each bridge exactly once?



Euler (1735): a path of the desired form is possible if and only if there are exactly zero or two nodes of odd degree.

Systems Biology in Graph-theoretic Terminology

- Instead of analyzing single small graphs and the properties of individual vertices or edges within such graphs, consider large-scale properties of graphs.
- This is now possible due to the availability of large amounts of data.

Why Graph-theoretic Properties?

- Large amounts of data, and hence manual analysis is not possible

- Simple questions (“what node is central?”) are replaced by more relevant questions (“what percent of nodes is crucial to the network connectivity?”)

- “How can I tell what this network looks like, even when I can’t actually look at it?”

Goals of This Lecture

- Statistical properties that characterize the structure and behavior of networked systems, and appropriate ways to measure these properties

- Models of networks that help understand the meaning of these properties

- Predict the behavior of networked systems based on measured structural properties and the local rules governing individual vertices

Lecture Outline

- Types of networks
- Networks in the real world
- Properties of networks
- Random graphs
- Exponential random graphs and Markov graphs
- The small-world model
- Models of network growth

Part I

Part II

Part I
Types and Properties of Networks

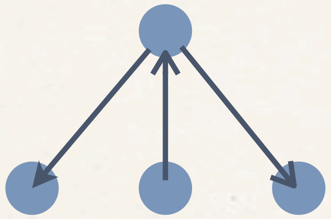
Types of Networks

- A “set of nodes joined by edges” is the simplest network model
- There may be types of nodes, types of edges; nodes and edges may have properties; nodes and edges can carry weights; edges can be directed

Example: Social Network of People

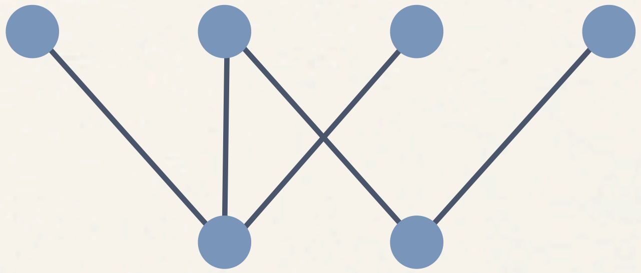
- Nodes: people
  - Different types: men, women, nationality, location, age, income,...
- Edges: relationships
  - Different types: friendship, animosity, professional acquaintance, geographical proximity,...
  - Weight of an edge: how well two people know each other,...

Special Types of Graphs



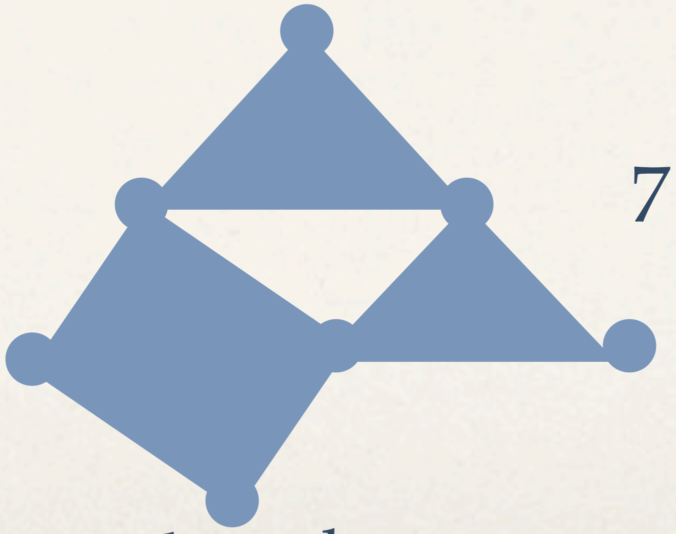
Digraph
(directed graph)

Edges are directed and paths can be traversed only along the direction of the edges



Bipartite graph

Two types of nodes, and edges connect between nodes of different types



7 nodes and 3 hyperedges

Hypergraph: contain edges that join more than two nodes

Networks in the Real World

Social Networks

- A social network is a set of people or groups of people with some pattern of contacts or interactions between them.

- Traditional social network studies (friendships, patterns of sexual contacts, etc.) often suffer from problems of inaccuracy, subjectivity, and small sample size.

- More reliable data in this area: collaboration networks, communication records (phone calls, emails, etc.)

Information Networks

- Also known as “knowledge networks”
- Classic example: network of citations between academic papers
- Edges are directed
- There are social aspects to the citation patterns of papers too...
- Citation networks are acyclic (with very rare exceptions)

Information Networks

- Another example of information network: the World Wide Web (WWW)
- It is cyclic
- Our picture of the network structure of WWW is biased: data comes from “crawls” of the network, in which Web pages are found by following hyperlinks from other pages
- There are social aspects to it, as well

Technological networks

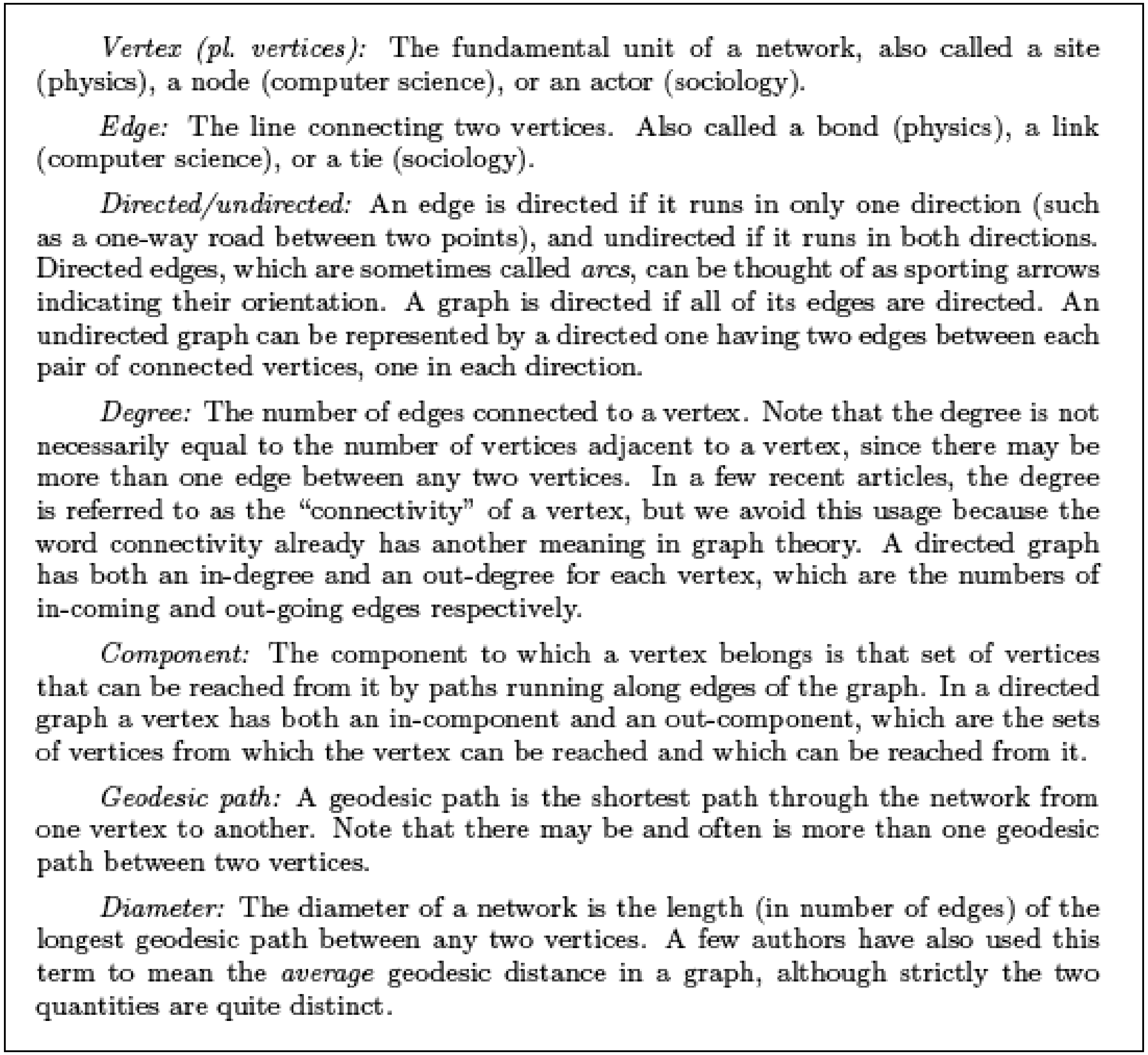
- Man-made networks designed for distribution of some commodity or resource, such as electricity or information
- Examples: the electric power grid, network of airline routes, network of roads, ...
- Another example: the Internet (the network of physical connections between computers)

Biological Networks

- Examples: network of protein interactions, metabolic pathways, signal transduction pathways, the food web, ...

Properties of Networks

Glossary of Terms



Vertex (pl. vertices): The fundamental unit of a network, also called a site (physics), a node (computer science), or an actor (sociology).

Edge: The line connecting two vertices. Also called a bond (physics), a link (computer science), or a tie (sociology).

Directed/undirected: An edge is directed if it runs in only one direction (such as a one-way road between two points), and undirected if it runs in both directions. Directed edges, which are sometimes called *arcs*, can be thought of as sporting arrows indicating their orientation. A graph is directed if all of its edges are directed. An undirected graph can be represented by a directed one having two edges between each pair of connected vertices, one in each direction.

Degree: The number of edges connected to a vertex. Note that the degree is not necessarily equal to the number of vertices adjacent to a vertex, since there may be more than one edge between any two vertices. In a few recent articles, the degree is referred to as the "connectivity" of a vertex, but we avoid this usage because the word connectivity already has another meaning in graph theory. A directed graph has both an in-degree and an out-degree for each vertex, which are the numbers of in-coming and out-going edges respectively.

Component: The component to which a vertex belongs is that set of vertices that can be reached from it by paths running along edges of the graph. In a directed graph a vertex has both an in-component and an out-component, which are the sets of vertices from which the vertex can be reached and which can be reached from it.

Geodesic path: A geodesic path is the shortest path through the network from one vertex to another. Note that there may be and often is more than one geodesic path between two vertices.

Diameter: The diameter of a network is the length (in number of edges) of the longest geodesic path between any two vertices. A few authors have also used this term to mean the *average* geodesic distance in a graph, although strictly the two quantities are quite distinct.

Properties of Networks: The small-world effect

- In a famous experiment carried out by S. Milgram in the 1960s, participants were asked to pass a letter to one of their first-name acquaintances in an attempt to get it to an assigned target individual.

- Result: the letters were able to reach the designated target individual in a small number of steps (around six in the published cases)

- The experiment demonstrated the *small-world effect*: most pairs of vertices are connected by a short path through the network.

Properties of Networks: The small-world effect

- The *mean geodesic (i.e., shortest) distance* between vertex pairs in a network with n nodes is

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}$$

the number of node pairs in the network

geodesic distance between nodes i and j

Computable in O(*mn*) time, where *m* is the number of edges, and *n* is the number of nodes

Properties of Networks: The small-world effect

- The mean geodesic distance is problematic when the network contains more than one component ($d_{ij}=\infty$ for some pairs in this case).
- One approach to overcome the problem: pairs of nodes that fall in two different components are excluded from the average
- An alternative, yet not as commonly used, approach uses the *harmonic mean* ($d_{ij}=\infty$ contribute nothing in this case), aka *graph efficiency*

$$\ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}$$

Properties of Networks: The small-world effect

- The small-world effect has obvious implications for the dynamics of processes taking place on networks.

- The small-world effect is mathematically obvious: if the number of vertices within a distance r of a typical central node grows exponentially with r (which is true for many networks, including random graphs), then the value of ℓ will increase as log n.

- Mathematically: networks are said to show the small-world effect if the value ℓ scales logarithmically or slower with network size for fixed mean degree

Properties of Networks: Transitivity (or, Clustering)

- “The friend of your friend is likely also to be your friend.”
- In network terminology: If vertex A is connected to vertex B, and vertex B is connected to vertex C, then there is a heightened probability that vertex A is also connected to vertex C.

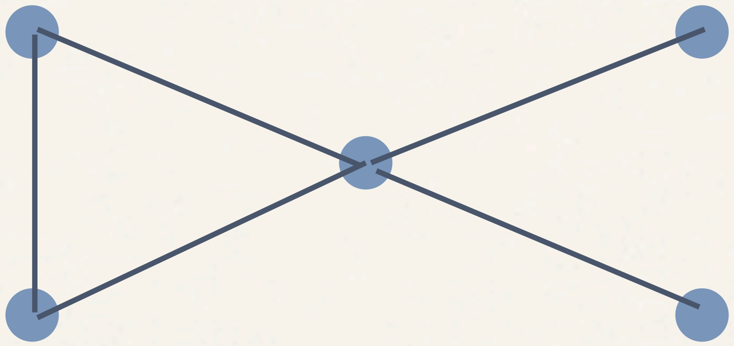
Properties of Networks: Transitivity (or, Clustering)

- In terms of network topology, transitivity implies the presence of a heightened number of triangles in the network
- The *clustering coefficient* is

$$C = \frac{3\times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}} \qquad (*)$$

Connected triple = a single vertex with edges running to an unordered pair of other vertices

Properties of Networks: An Example of Clustering Coefficient



In this network, there are:
1 triangle, and 8 connected triples

Therefore, based on Formula (*), C= 3/8

Properties of Networks: Clustering Coefficient

- Satisfies $0 \leq C \leq 1$

- C can be interpreted as the mean probability that two vertices that are network neighbors of the same other vertex will themselves be neighbors.

Properties of Networks:

Clustering Coefficient: Another Definition

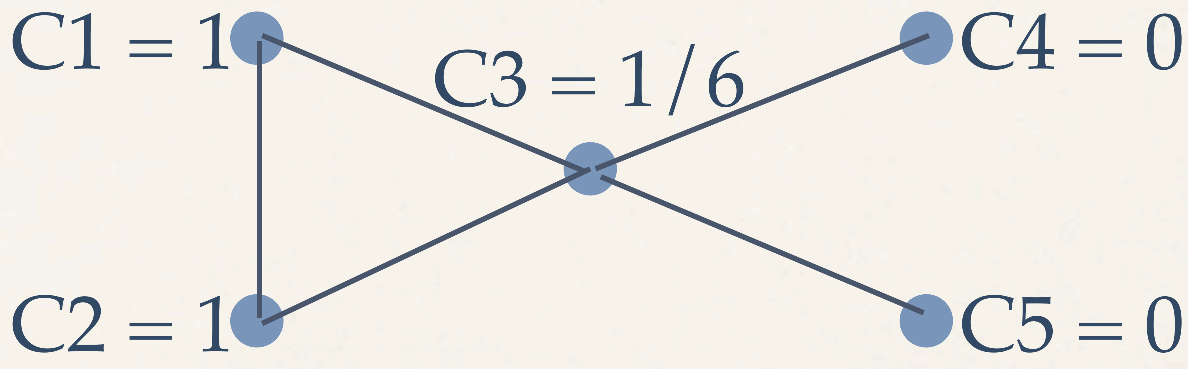
$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}$$

(C_i=0 if denominator is 0)

The clustering coefficient of the whole network

$$C = \frac{1}{n}\sum_i C_i \qquad (**)$$

Properties of Networks:
An Example of Clustering Coefficient



Based on Formula (**), C= 13/30

Properties of Networks:

Clustering Coefficient: Another Definition

- The clustering coefficient we described measures the density of triangles in a network.

- A number of authors have also looked at the density of longer loops, but no clean theory that separates the independent contributions of the various orders from one another

Properties of Networks: Node Degrees

k_i : degree of node i

For directed graphs: k_i^{in} and k_i^{out}

Average degree:

$$\langle k \rangle = \frac{1}{n}\sum_{i=1}^{n} k_i \qquad \langle k_i^{in} \rangle = \frac{1}{n}\sum_{i=1}^{n} k_i^{in} \qquad \langle k_i^{out} \rangle = \frac{1}{n}\sum_{i=1}^{n} k_i^{out}$$

Properties of Networks: Degree distributions

- Define p_k to be the fraction of vertices in the network that have degree k
- The degree distribution of a network can be visualized by making a histogram of the p_k values
- For (Erdos-Renyi) random graphs, the degree distribution is binomial
- Real-world networks have degree distributions with a very long right tail

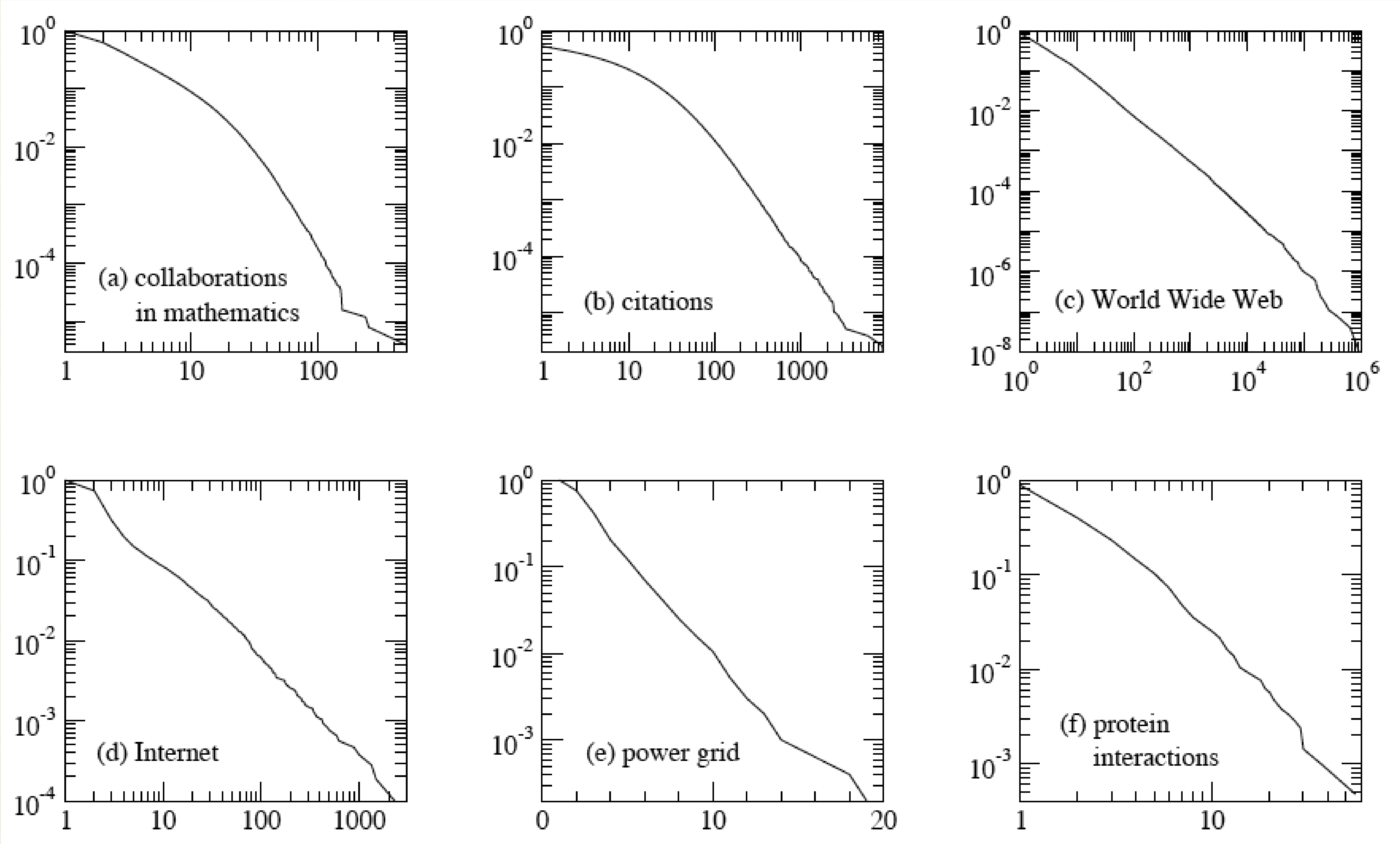
Properties of Networks: Degree distributions

- Measuring the tail is tricky: in practice there are rarely enough measurements to get good statistics in the tail
- Two approaches around this problem:
  - Construct a histogram in which the bin sizes increase exponentially with degree (1, 2-3, 4-7, 8-15,...)
  - Plot the cumulative distribution function

$$P_k = \sum_{k'=k}^{\infty} p_{k'}$$

Properties of Networks:

Cumulative Degree distributions for Six Different Networks



Properties of Networks: Power Law Degree Distribution

- As the figure on the previous slide shows, the distributions are all right-skewed.

- Many of them have power-law degree distributions: $p_k \sim k^{-\alpha}$

- They show up as power laws in the cdf, but with a smaller exponent:

$$P_k \sim \sum_{k'=k}^{\infty} k'^{-\alpha} \sim k^{-(\alpha-1)}$$

Properties of Networks: Power Law Degree Distribution

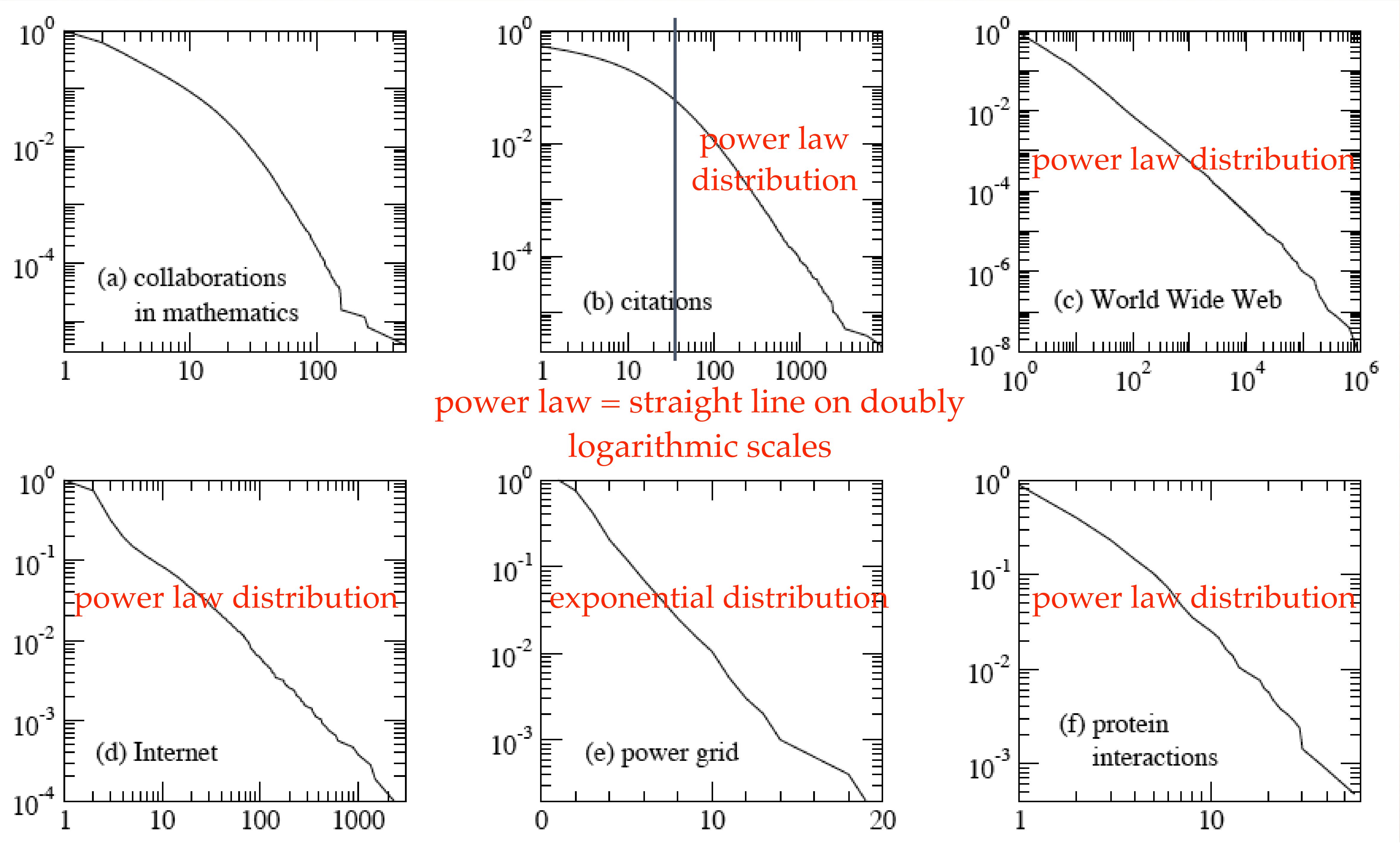
- Some of the other distributions (e.g., power grid) have exponential tails: $p_k \sim \mathrm{e}^{-k/\kappa}$

- They also give exponentials in the cdf, but with the same exponent:

$$P_k = \sum_{k'=k}^{\infty} p_k \sim \sum_{k'=k}^{\infty} \mathrm{e}^{-k'/\kappa} \sim \mathrm{e}^{-k/\kappa}$$

Properties of Networks:

Cumulative Degree distributions for Six Different Networks



power law distribution

power law distribution

power law = straight line on doubly logarithmic scales

power law distribution

exponential distribution

power law distribution

Properties of Networks:
Degree Distribution for Other Types of Networks

- Bipartite graphs: two distributions, one for each type of nodes
- Directed graphs: nodes have in- and out-degrees, and the degree distribution therefore is a function of two variables

$$p_{jk}$$

the fraction of nodes that simultaneously have in-degree j and out-degree k

Properties of Networks: Scale-free Networks

- Networks with power-law degree distributions are also referred to as *scale-free networks*.

- Scale-free refers to any functional form f(x) that remains unchanged to within a multiplicative factor under a rescaling of x

- This, in effect, means power-law forms, since these are the only solutions to f(ax)=bf(x).

Properties of Networks: Scale-free Real Life Networks

- Citation networks
- WWW
- The Internet
- Metabolic networks
- Telephone call graphs
- Network of human sexual contacts

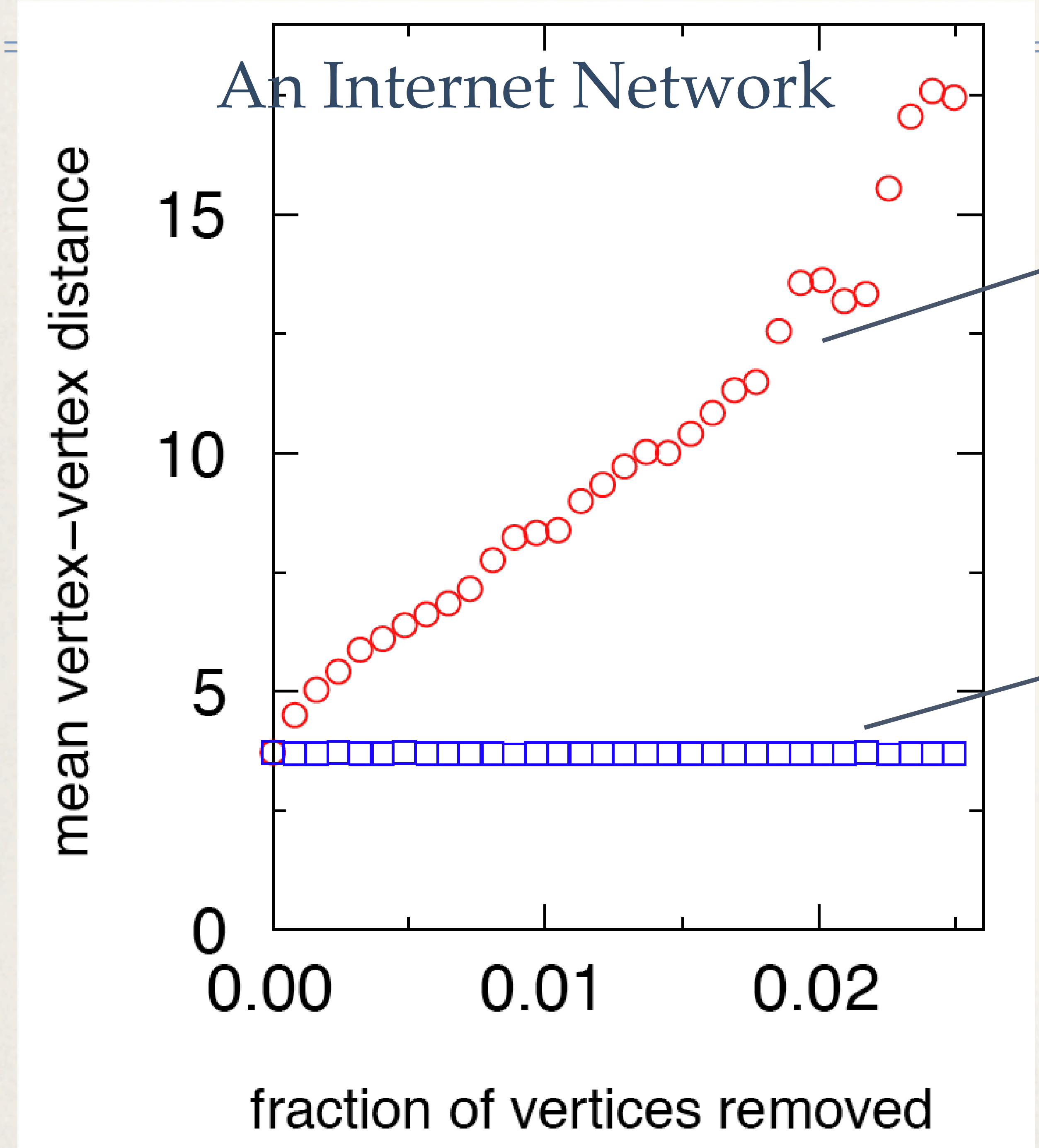
Properties of Networks:
Real Life Networks with Other Degree Distribution

- Exponential
  - Power grid
  - Railway networks
- Power law with exponential cutoffs
  - Network of movie actors
  - Some collaboration networks

Properties of Networks: Network resilience

- Resilience of networks to the removal of their vertices is important.
- As vertices are removed, the geodesic distances increase, and eventually become infinite (i.e., the network becomes disconnected).
- Networks vary in their level of resilience to vertex removal.
- An example from epidemiology: vaccination of individuals is modeled by the removal of their corresponding vertices, and the effect of such removal is of interest (e.g., what effect it has on the spread of the disease)

Properties of Networks: Network resilience: An Example

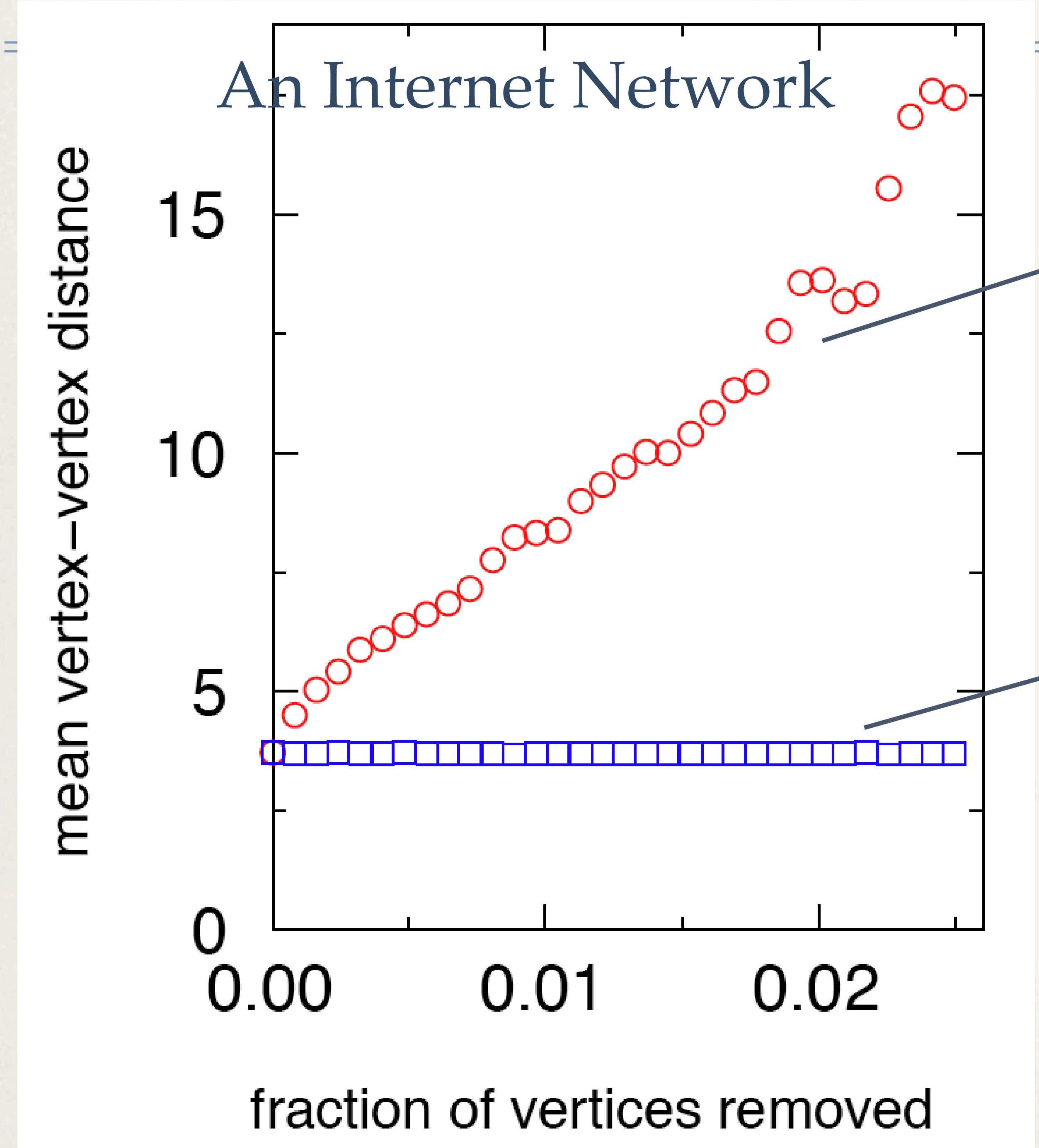


An Internet Network

Vertices removed in decreasing order of their degree

Vertices removed in random order

Properties of Networks: Network resilience: An Example



An Internet Network

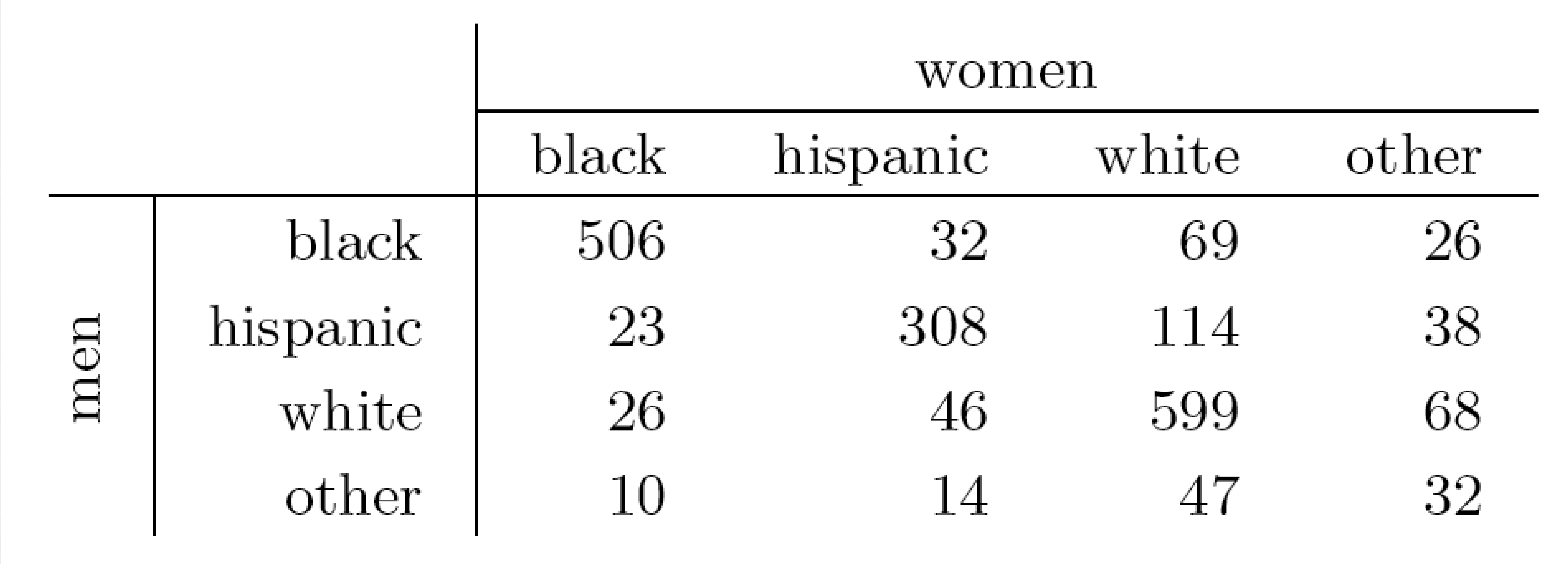
Vertices removed in decreasing order of their degree

Vertices removed in random order

Conclusion: The Internet is highly resilient to random failures, but highly vulnerable to deliberate attacks on highest-connectivity nodes

Properties of Networks: Mixing patterns

- In a network with multiple types of nodes, how do types affect connectivity? Do nodes associate with others from the same type (*assortative*), or from different types (*disassortative*)?



		women			
		black	hispanic	white	other
men	black	506	32	69	26
	hispanic	23	308	114	38
	white	26	46	599	68
	other	10	14	47	32

Mixing by race in a social network

Properties of Networks: Mixing patterns

- Assortative mixing can be quantified by an *assortativity coefficient*
- E_{ij}: the number of edges connecting nodes of types i and j
- **E**: the matrix with elements E_{ij}
- e=**E**/sum(**E**): the normalized mixing matrix (sum(**E**) is the sum of all elements in **E**)

Properties of Networks: Mixing patterns

- Another measure is the conditional probability that vertex x's neighbor is of type j given that x is of type i

$$P(j|i) = \frac{e_{ij}}{\sum_k e_{ik}}$$

Properties of Networks:

Mixing patterns: Assortativity coefficients

$$Q = \frac{\sum_i P(i|i) - 1}{N - 1}$$

(N: number of types)

Q has two problems: (1) may give different values for asymmetric input matrices, and (2) weights all types equally.

A measure that remedies both problems:

$$r = \frac{\mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|}{1 - \| \mathbf{e}^2 \|}$$

(Tr: trace)

Properties of Networks: Degree correlations

- Do high-degree vertices associate preferentially with other high-degree vertices or do they prefer to attach to low-degree ones?

- One way to measure degree correlations in a network is to compute *Pearson correlation coefficient* of the degrees at either ends of an edge.

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{(m-1)\sigma_x \sigma_y}$$

m : the number of edges; (x_i, y_i): the degrees of endpoints of edge i

$(\overline{x}, \overline{y})$: the means, and σ_x, σ_y : the standard deviations

Properties of Networks: Community structure

- Are there communities within the network such that there is a high density of edges within the community, but not between communities?

- The traditional method for extracting community structure is *cluster analysis* (or, hierarchical clustering)

- Connection strength is assigned to each vertex pair, and then, starting from n vertices with no edges, add edges between pairs of vertices in decreasing order of connection strength, and observe the emerging structure

Properties of Networks: Community structure

- Various ways for defining connection strength: weighted vertex-vertex distance measures, sizes of minimum cut-sets, ...
- Another way is to measure *edge betweenness*: the number of geodesic paths between vertices that run along each edges in the network.
- Under this measure, biological networks seem to exhibit community structure.

Properties of Networks: Network navigation

- Milgram’s famous small-world experiment showed that there are short paths between apparently distant individuals.
- A more surprising result is that the experiment showed that “ordinary people were good at finding these short paths”, even though the participants had no knowledge of the structure of the network (they only knew their “neighbors”).
- In random networks, on the other hand, short paths exist, but no one would be able to find them given the kind of information that people have in realistic scenarios

Properties of Networks: Other Network Properties

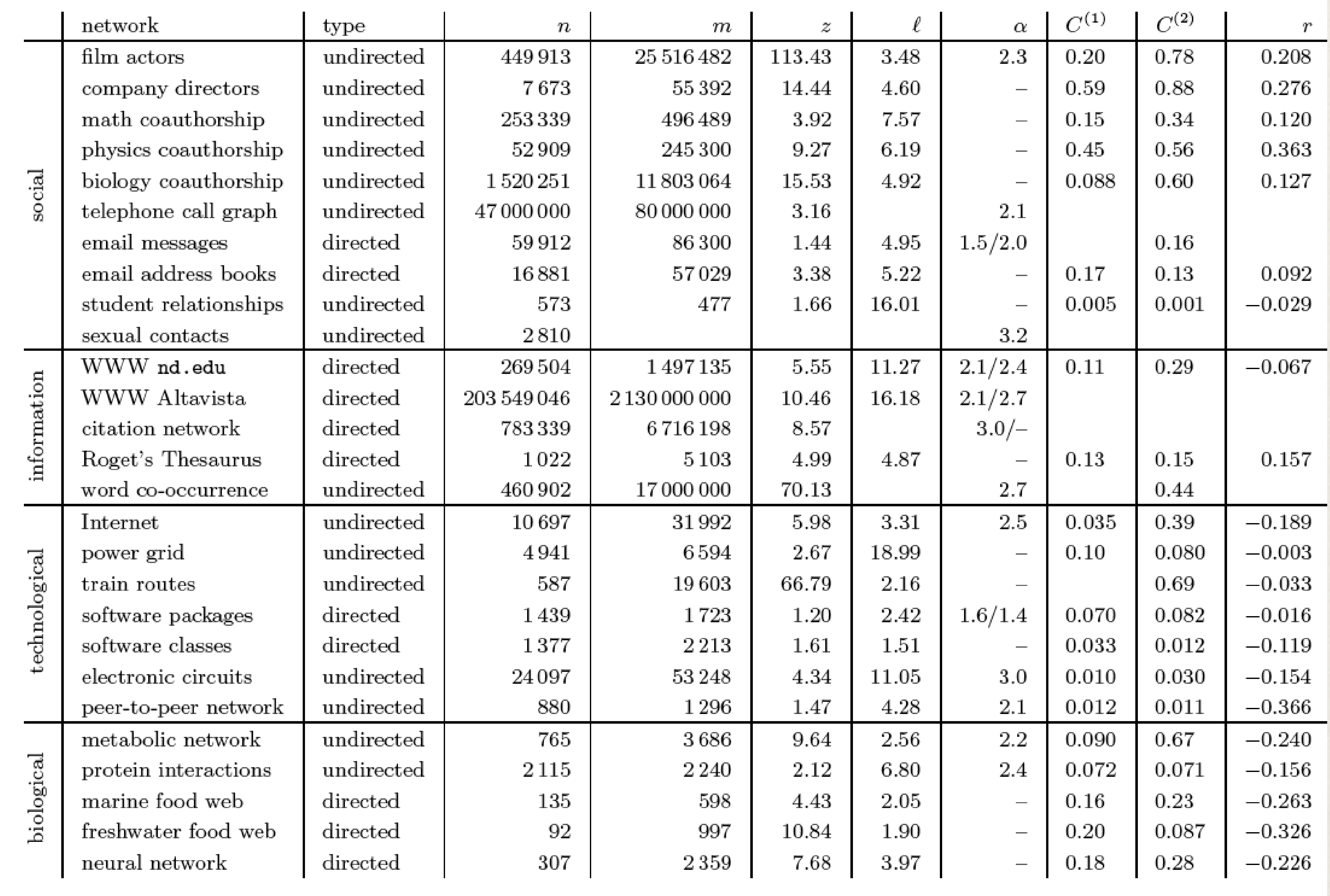
- People have looked at the size of the largest component, second largest component, etc., in a network

- *Betweenness centrality* of a vertex v: the number of geodesic paths between other vertices that run through v

- Betweenness centrality can be viewed as a measure of network resilience

- *Efficiency* of a vertex v: the harmonic mean distance between v and all other vertices

- *Eigenvalues* and *eigenvectors* of adjacency matrix of the network; *network motifs*; and many more...



	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1			
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16	
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029
	sexual contacts	undirected	2 810				3.2			
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7			
	citation network	directed	783 339	6 716 198	8.57		3.0/–			
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44	
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226

Part II
Models of Networks

Why Network Models?

- Take observed properties of real-world networks, and try to create models that would generate networks with similar properties
- This leads to a better understanding of real-world network and the ability to simulate and analyze them, as well as predict behavior of processes that take place on them.

Random Graphs:

The Erdős-Rényi (ER) model (Poisson random graphs)

- To generate a random number on n nodes: (1) take n nodes, and (2) connect each pair with probability p
- $G_{n,p}$ is the ensemble of all such graphs
- A graph in $G_{n,p}$ has probability $p^m(1-p)^{M-m}$ of having m edges, where $M = \frac{1}{2}n(n-1)$

Random Graphs: Properties of the ER Model

It has Poisson degree distribution with mean $z = p(n-1)$

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{z^k e^{-z}}{k!}$$

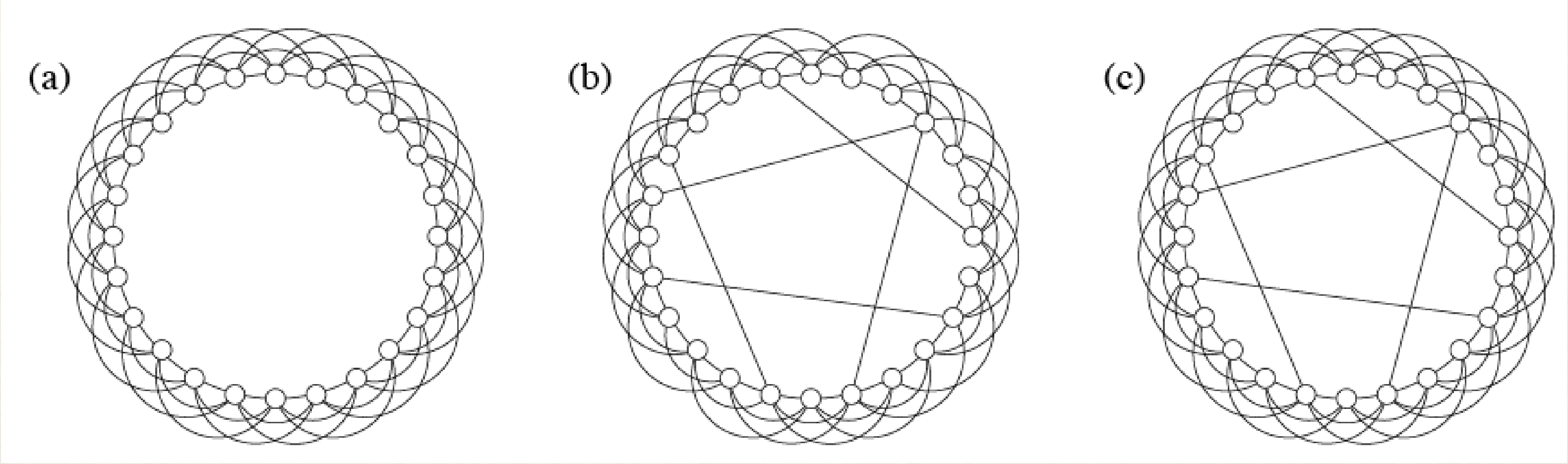
Has the small-world effect

Has a low clustering coefficient C=p, since the probability of connection of two vertices is p, regardless of whether they have a common neighbor.

It has entirely random mixing patterns; no correlation between degrees of adjacent vertices; no community structure, and navigation is impossible using local algorithms.

The Small-world Model

- Small-world models have high degree of transitivity
- They are built on lattices of any dimension or topology
- Creating a small-world network on a one-dimensional lattice:
  - Take a ring with L vertices
  - Join each vertex to its neighbors k or fewer lattice spacings away
  - For each edge, and with probability p, move one end of that edge to a new location chosen uniformly at random from the lattice, except that no double edges or self-edges are ever created (in a slightly modified version, no edges are removed from the underlying lattice)



The Small-world Model: Properties

- A regular lattice (p=0) does not show the small-world effect: mean geodesic distances between vertices tend to L/4k for large L, and have clustering coefficient of C=(3k-3)/(4k-2)
- For the small-world model, Barrat and Weigt showed that $C = \frac{3(k-1)}{2(2k-1)}(1-p)^3$
- For the modified version, Newman showed that $C = \frac{3(k-1)}{2(2k-1)+4kp(p+2)}$

Models of Network Growth

- Models we've seen so far do not explain how networks come to have their properties

- In this part, we will discuss models that reflect how networks grow, and how growth leads to the networks' properties

Models of Network Growth: Price's Model: The Rationale

- Price called his model *cumulative advantage*, and is today known under the name *preferential attachment* (coined by Barabási and Albert)

- Originally it was done in the context of paper citations, and Price's idea was that *the rate at which a paper gets new citations should be proportional to the number that it already has*

- Makes sense: the more a paper is cited, the higher the probability that you'll come across it while reading the literature, and cite it

Models of Network Growth: Price's Model: The Details

- Consider a directed graph of n vertices
- Let p_k be the fraction of vertices with in-degree k
- New vertices are continually added, not necessarily at a constant rate
- Each added vertex has a certain out-degree, which is fixed permanently at the creation of the vertex
- The out-degree may vary from vertex to another, but the mean degree, denoted by m, is a constant over time
- Price suggests that the probability of attaching a new edge to a vertex of in-degree k is proportional to $k+k_0$, where k_0 is a constant (he chose $k_0=1$, which we also use)

Models of Network Growth: Price's Model: The Details

The probability that a new edge attaches to any of the vertices with degree k is

$$\frac{(k+1)p_k}{\sum_k (k+1)p_k} = \frac{(k+1)p_k}{m+1} \qquad (1)$$

The mean number of new edges to vertices with current in-degree k is

$$\frac{(k+1)p_k m}{m+1} \qquad (2)$$

The number of vertices of in-degree k, np_k, decreases by the amount in (2), but increases because of the influx from the vertices previously of degree k-1 that have just acquired a new edge (except for vertices of degree 0, which have an influx of exactly 1)

Models of Network Growth: Price's Model: The Details

Denote by $p_{k,n}$ the value of p_k when the graph has n vertices, then the net change in np_k per vertex added, for k>0, is

$$(n+1)p_{k,n+1} - np_{k,n} = [kp_{k-1,n} - (k+1)p_{k,n}]\frac{m}{m+1}$$

and, for k=0, is

$$(n+1)p_{0,n+1} - np_{0,n} = 1 - p_{0,n}\frac{m}{m+1}$$

Looking for stationary solutions p_k,n+1=p_k,n=p_k, we find:

$$p_k = \begin{cases} [kp_{k-1} - (k+1)p_k]m/(m+1) & \text{for } k \geq 1 \\ 1 - p_0 m/(m+1) & \text{for } k = 0 \end{cases} \qquad (3)$$

Rearranging (3), we find

$$p_0 = (m+1)/(2m+1) \qquad\qquad p_k = p_{k-1}k/(k+2+1/m)$$

Models of Network Growth: Price's Model: The Details

The rearrangement of (3) can be written as

$$p_k = \frac{k(k-1)\cdots 1}{(k+2+1/m)\cdots(3+1/m)} p_0 = (1+1/m)B(k+1, 2+1/m)$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is Legendre's beta-function, which goes asymptotically as a^{-b} for large a and fixed b, and hence

$$p_k \sim k^{-(2+1/m)}$$

In other words, in the limit of large n, the degree distribution of a network growing according to Price's model has a power-law tail with exponent

$$\alpha = 2 + 1/m$$

This will typically give exponents in the interval between 2 and 3, which is in agreement with the values seen in real-world networks.

Models of Network Growth: Price's Model: The Details

For the generalized case of $k_0 \neq 1$ we have

$$p_k = \frac{m+1}{m(k_0+1)+1} \frac{B(k+k_0, 2+1/m)}{B(k_0, 2+1/m)}$$

and hence, for large k and fixed k_0, we have

$$\alpha = 2 + 1/m$$

Models of Network Growth: The Barabási-Albert (BA) model

- The BA model is the same as Price's with one important difference: in the BA model, the edges that are added are undirected, so there is no distinction between in- and out-degree.

- Each vertex in the network appears with initial degree m, which is never changed thereafter, and the other end of each edge being attached to another vertex with probability proportional to the degree of that vertex.

Models of Network Growth: The BA Model: The Details

The probability that a new edge attached to any vertex of the vertices with degree k is

$$\frac{kp_k}{\sum_k kp_k} = \frac{kp_k}{2m} \qquad (1)$$

The mean number of new edges to vertices with current degree k is

$$m \times kp_k/(2m) = \frac{1}{2}kp_k \qquad (2)$$

The number of vertices of degree k, np_k, decreases by the amount in (2), but increases because of the influx from the vertices previously of degree k-1 that have just acquired a new edge (except for vertices of degree m, which have an influx of exactly 1)

Models of Network Growth: The BA Model: The Details

Denote by $p_{k,n}$ the value of p_k when the graph has n vertices, then the net change in np_k per vertex added, for k>m, is

$$(n+1)p_{k,n+1} - np_{k,n} = \frac{1}{2}(k-1)p_{k-1,n} - \frac{1}{2}kp_{k,n}$$

and, for k=m, is

$$(n+1)p_{m,n+1} - np_{m,n} = 1 - \frac{1}{2}mp_{m,n}$$

Looking for stationary solutions $p_{k,n+1}=p_{k,n}=p_k$, we find:

$$p_k = \begin{cases} \frac{1}{2}(k-1)p_{k-1} - \frac{1}{2}kp_k & \text{for } k > m \\ 1 - \frac{1}{2}mp_m & \text{for } k = m \end{cases} \qquad (3)$$

Rearranging (3), we find

$$p_m = 2/(m+2) \qquad\qquad p_k = p_{k-1}(k-1)/(k+2)$$

Models of Network Growth: The BA Model: The Details

The rearrangement of (3) can be written as

$$p_k = \frac{(k-1)(k-2)\cdots m}{(k+1)(k+1)\cdots(m+3)} p_m = \frac{2m(m+1)}{(k+2)(k+1)k}$$

In the limit of large k, this gives a power law degree distribution

$$p_k \sim k^{-3}$$

Two other interesting properties of the model:

1. a correlation between the age and degree of vertices: older vertices have higher mean degree
2. a correlation between the degrees of adjacent vertices in the model

Models of Network Growth: Generalizations of the BA model

- Several extensions of the BA model have been proposed:
  - Probability of attachment to a vertex is proportional to $k+k_0$, where k_0 is a constant that is allowed to be negative
  - The probability of attachment to a vertex is not linear in the degree of the vertex
  - The mean degree changes over time

Models of Network Growth: Vertex copying models

- Biochemical reaction networks appear to have power-law degree distributions, yet preferential attachment does not seem to be the appropriate model.

- For example, protein interaction networks (proteins are vertices and interactions are edges) change on a very long time-scale due to evolution, but biologically, cumulative advantage or preferential attachment are too simplistic to capture the evolutionary change.

- Kleinberg's model: the graph grows by stochastically constant addition of vertices and addition of directed edges by copying them from another vertex:

  - Choose an existing vertex u and a number m of edges to add to it

  - Choose a random vertex v and copy targets from m of its edges to u

  - If the chosen vertex v has less than m outgoing edges, then its m edges are copied and one moves to a new vertex and copies its edges, and so forth until m edges in total have been copied

Models of Network Growth: Vertex copying models

- Kleinberg’s model gives rise to power-law degree distributions

- The rationale for using this model in biological networks is based on gene duplication: the proteins encoded by the two copies of the gene have the same interactions upon duplication...

Acknowledgments

- Materials in this lecture are mostly based on:
  - "The Structure and Function of Complex Networks", by M.E.J. Newman, and references therein.