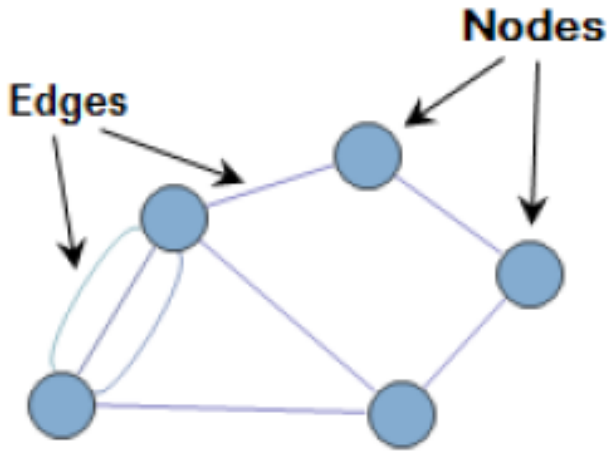
Network Structure

Michael Goodrich

Some slides adapted from:
Networked Life (NETS) 112, Univ. of Penn., 2018, Prof. Michael Kearns
Kentaro Toyama - Microsoft Research India

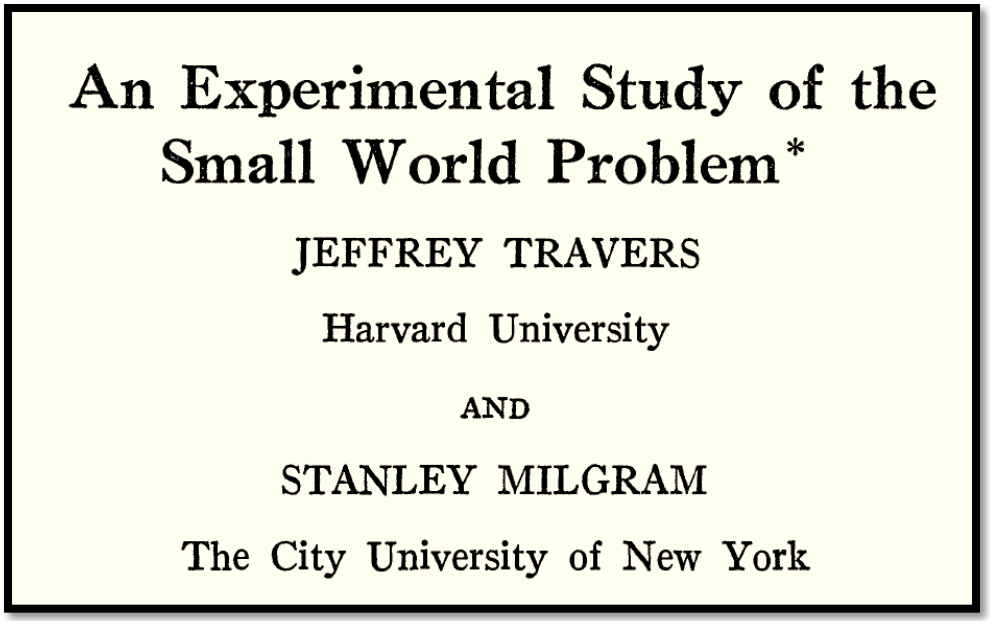
Terminology

- **Graph:** a network, G=(V,E), consists of two sets, V, of vertices and, E, edges.
  - **Vertices:** these are the entities in graph (also called nodes or actors). For example, if we consider Facebook friends as a graph, then every friend is a vertex.
  - **Edges:** These are pairwise relationships between vertices. For example, if we consider Facebook friends as a graph then every friendship is an edge. Edges are sometimes also called “ties.”



The Degrees-of-Separation Experiment

- An experiment by Travers and Milgram published in 1969 to determine how many acquaintance “hops” exist between people in America.



An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

The City University of New York

The Degrees-of-Separation Experiment

- Random people in Nebraska were asked to send a letter to a stockbroker in Boston, MA.
- Participants could only mail the letter to personal acquaintances, sending it directly only if they knew the stockbroker and otherwise sending it to a friend they thought might be able to reach the target.

The Degrees-of-Separation Experiment

- 64 out of the 296 starting letters succeeded to be delivered to the stockbroker in Boston.
- The average number of hops for successful letters was between 5 and 6.

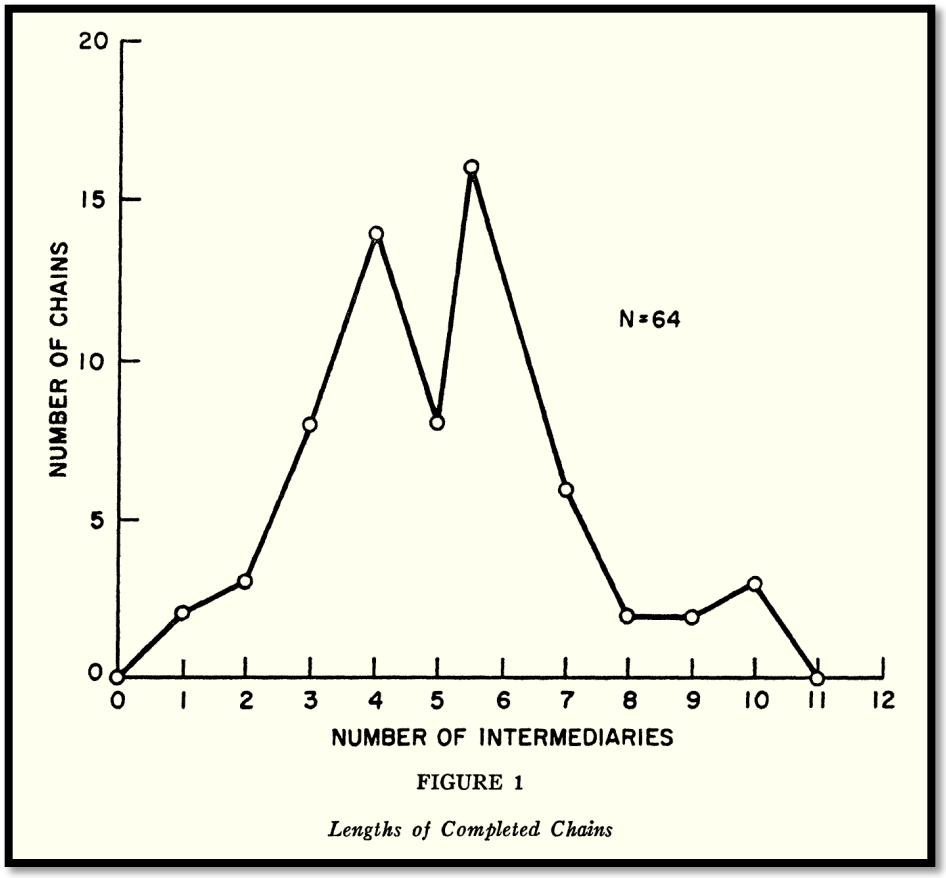
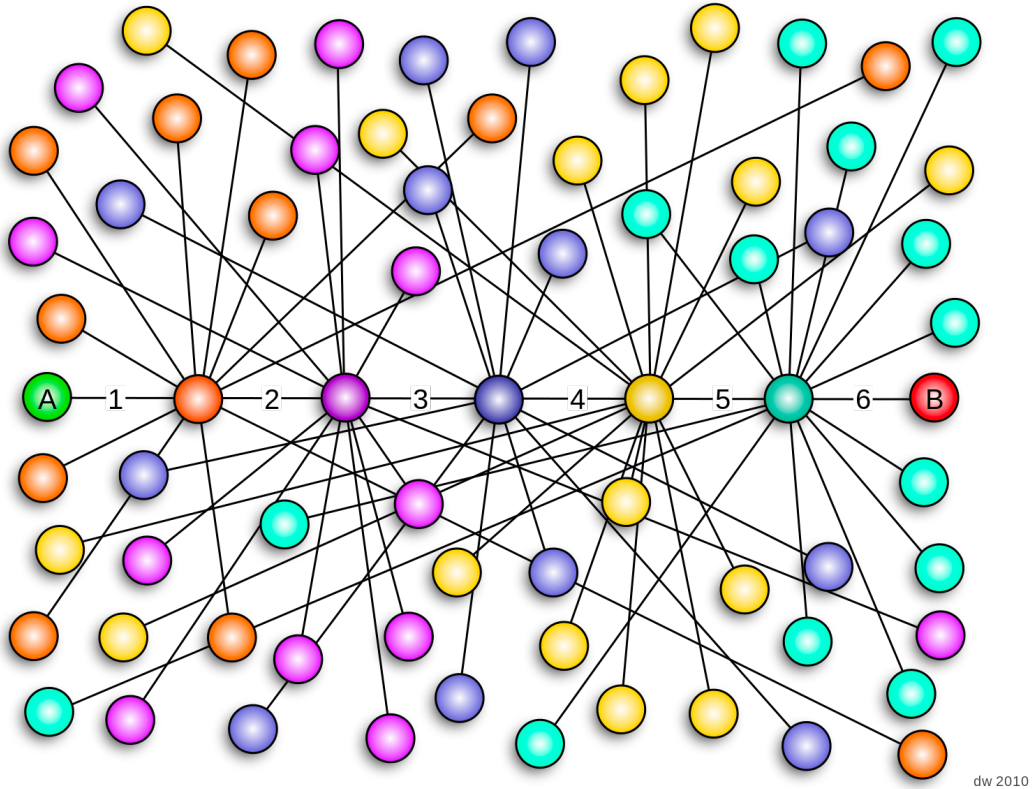


FIGURE 1

Lengths of Completed Chains

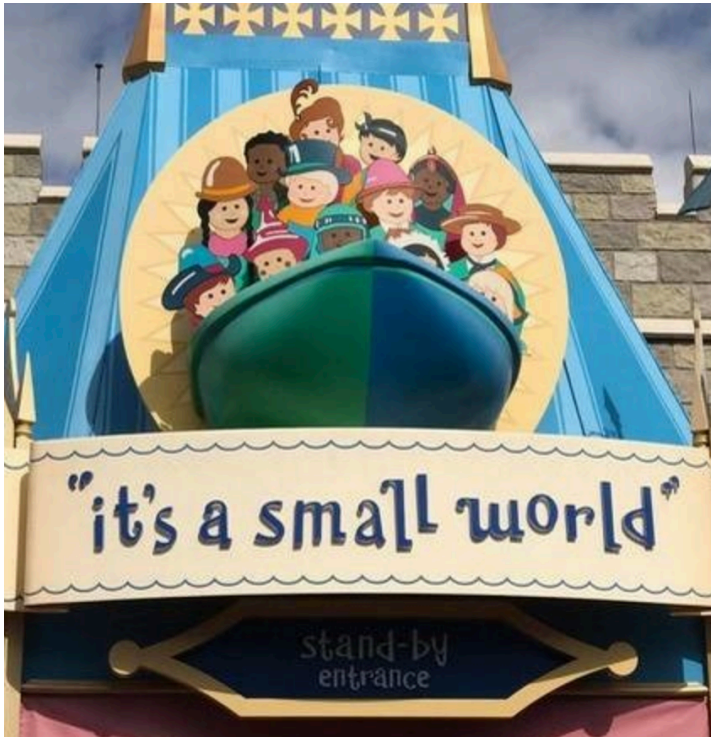
The Degrees-of-Separation Experiment

- This result gave rise to the expression that everyone in America has at most **six degrees of separation**.



Small Worlds

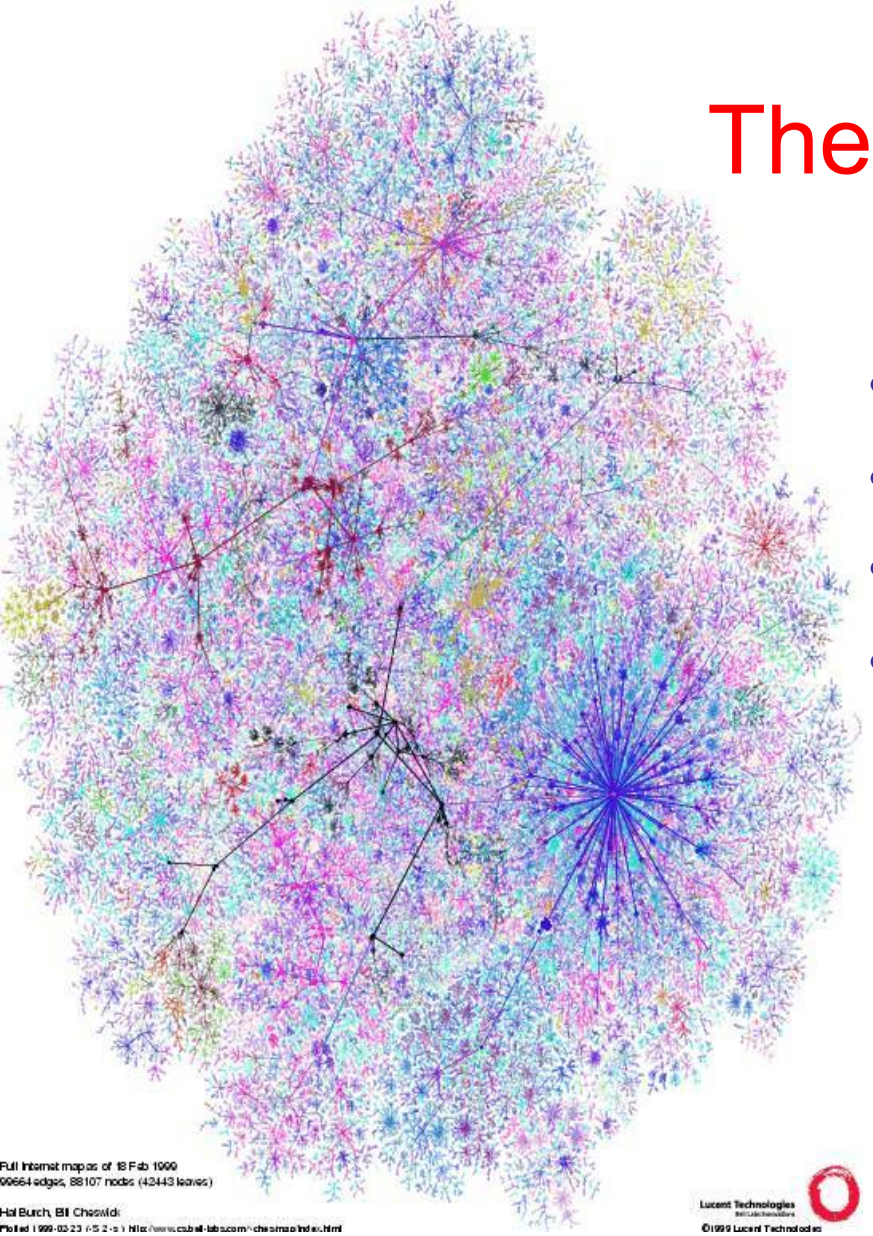
- This experiment led to many other, similar experiments and to a popular ride at Disneyland.*



*Actually, the Disneyland ride, "It's a Small World," opened in 1966, three years before the Travers-Milgram experiment was published.

Example Networks (Social and Otherwise)

The Internet, Router Level



- "Vertices" are physical machines
- "Edges" are physical wires
- Interaction is electronic
- A purely technological network?

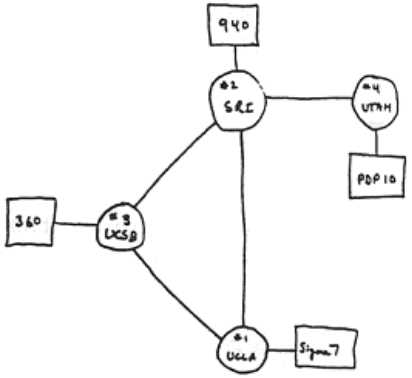
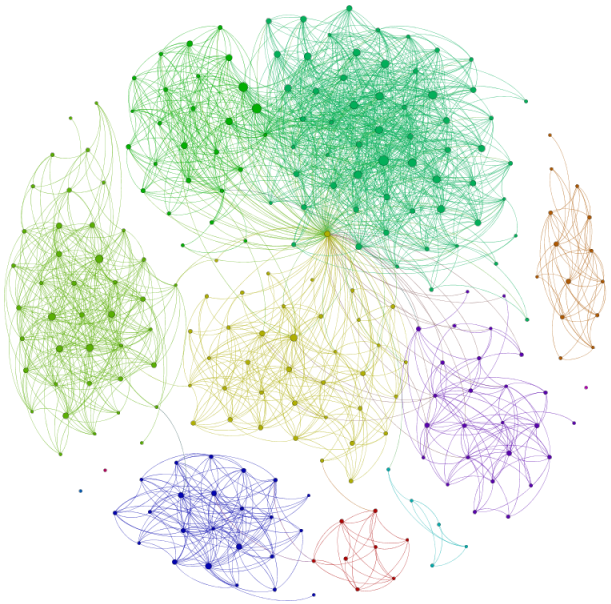
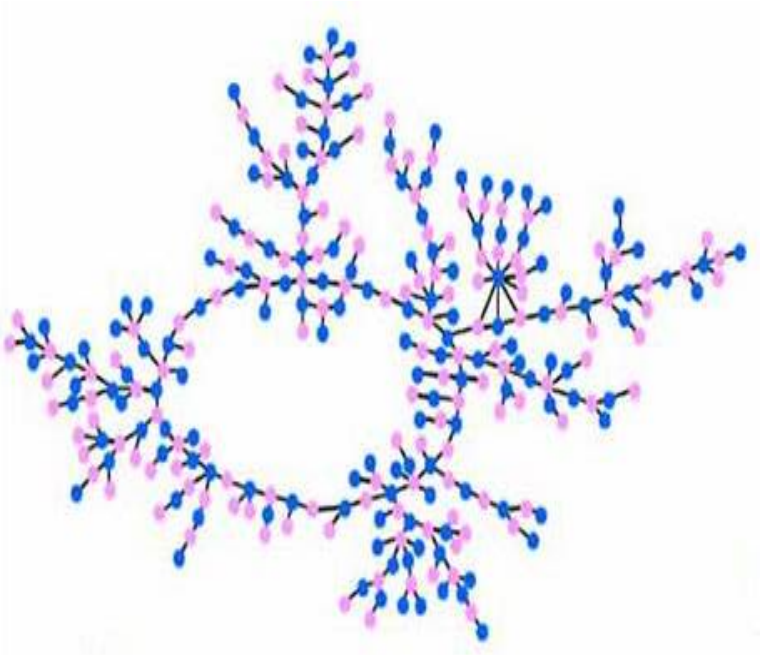


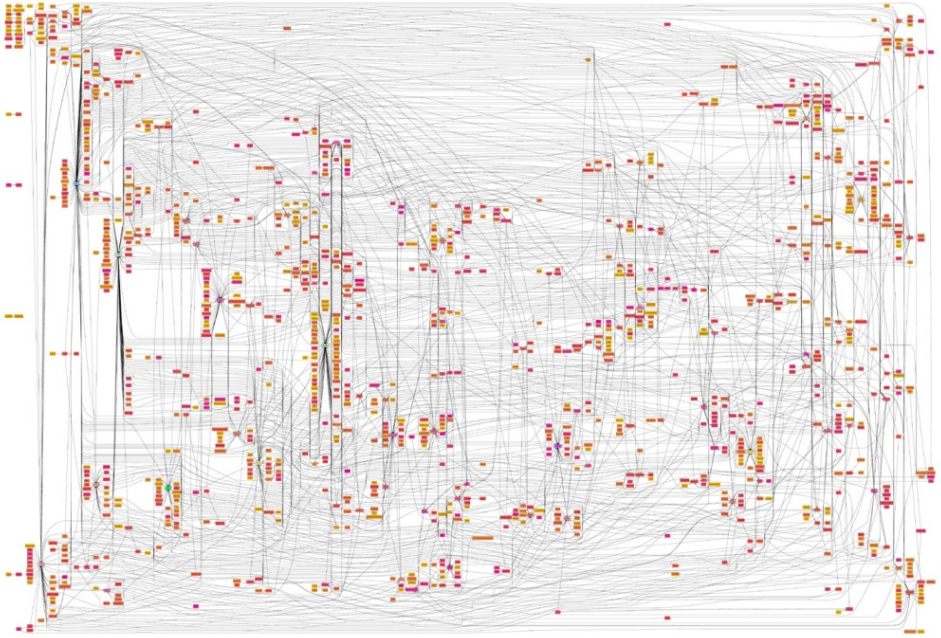
FIGURE 6.2 Drawing of 4 Node Network
(Courtesy of Alex McKenzie)



Social Networks

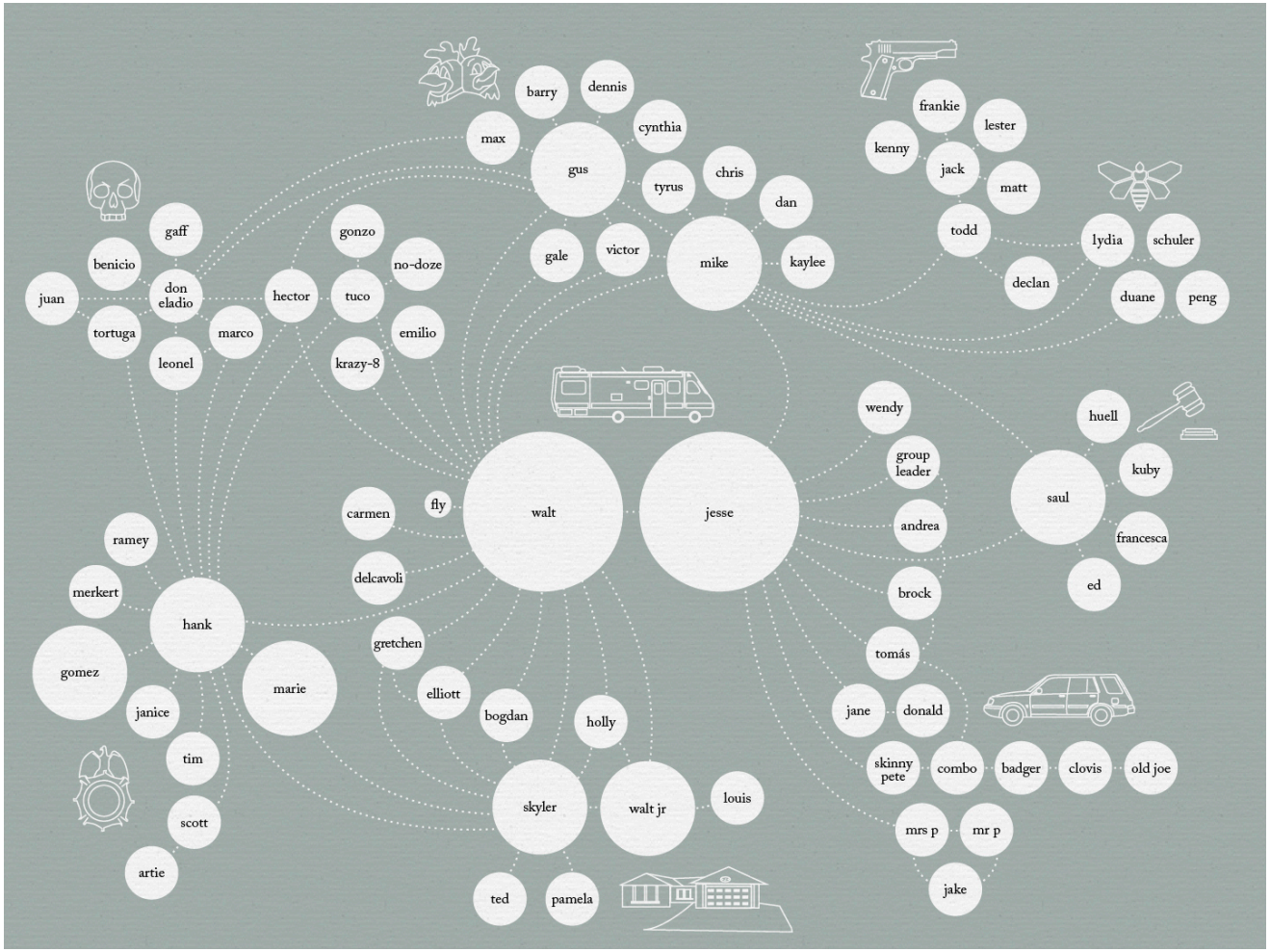
- Vertices are *people or animals*
- Edges are *social relationships*
- Interactions: relationships, professional, virtual…
- How and why does *structure* form?





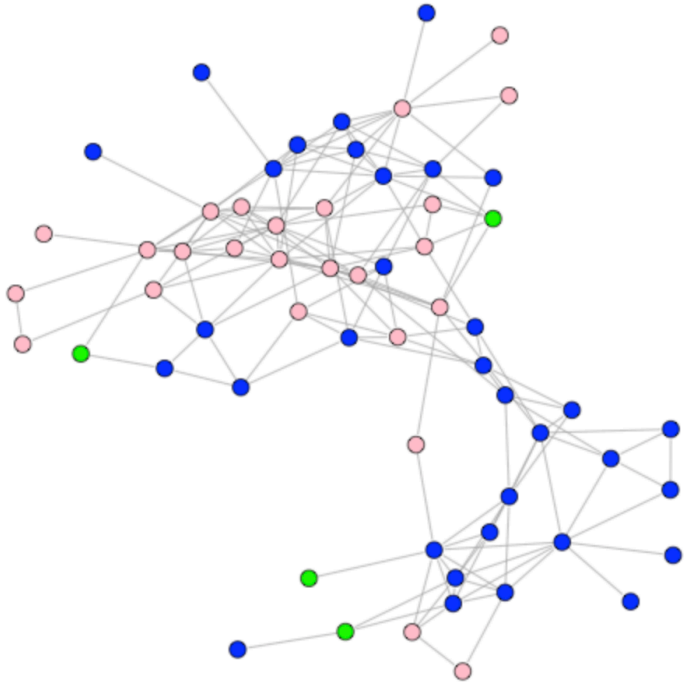
A Small Social Network

- From the TV show, “Breaking Bad”:



Another Small Social Network

- Bottlenose dolphins
  - Vertices are dolphins
  - Edges are observed companionships



The dolphin social network with sex: pink (female), blue (male), green (unknown). Most of the links (70%) connect dolphins of the same sex.

From https://users.dimi.uniud.it/~massimo.franceschet/bottlenose/bottlenose.html

Protein-Protein Interaction Networks

- Vertices are proteins in a cell.
- Edges represent the physical contacts between proteins in the cell.

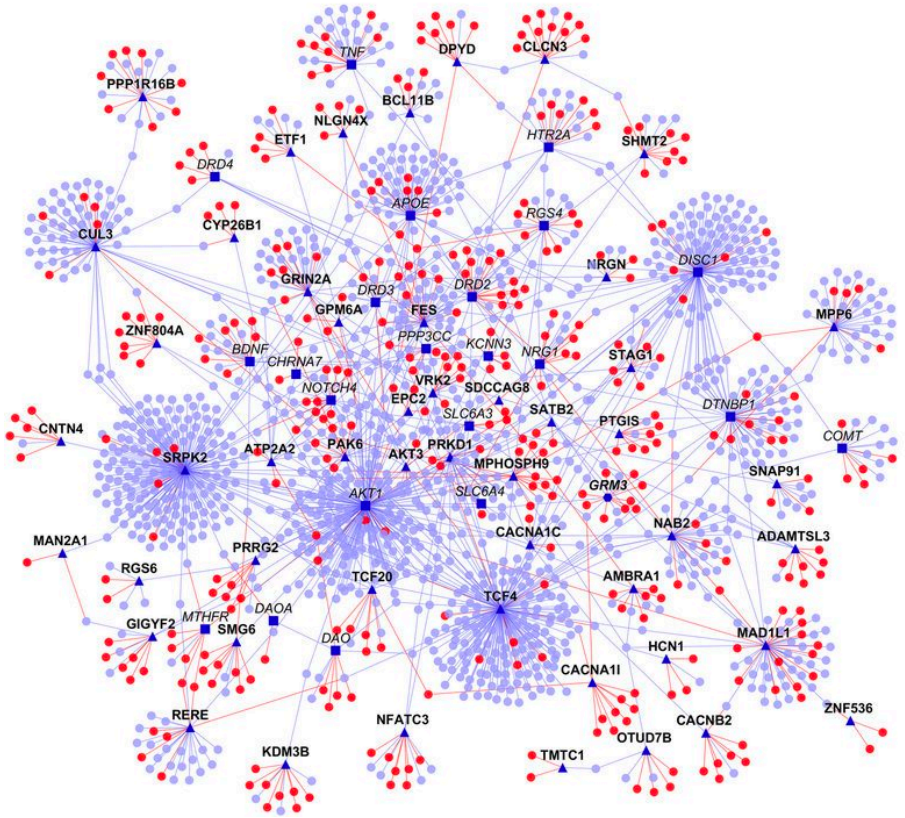


Image by Madhavicmu - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=48447204

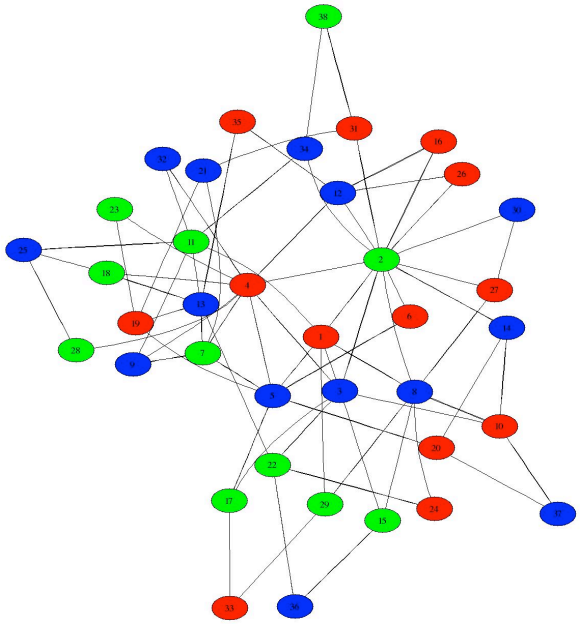
Who's Doing All This?

- Computer Scientists
  - Understand and design complex, distributed networks
  - View "competitive" decentralized systems as economies
- Social Scientists, Behavioral Psychologists, Economists
  - Understand human behavior in "simple" settings
  - Revised views of economic rationality in humans
  - Theories and measurement of social networks
- Biologists
  - Protein-protein interaction networks, gene regulatory networks,…
- Physicists and Mathematicians
  - Interest and methods in complex systems
  - Theories of macroscopic behavior (phase transitions)
- Communities are *interacting* and *collaborating*

Structural Properties of Networks

Networks: Basic Definitions

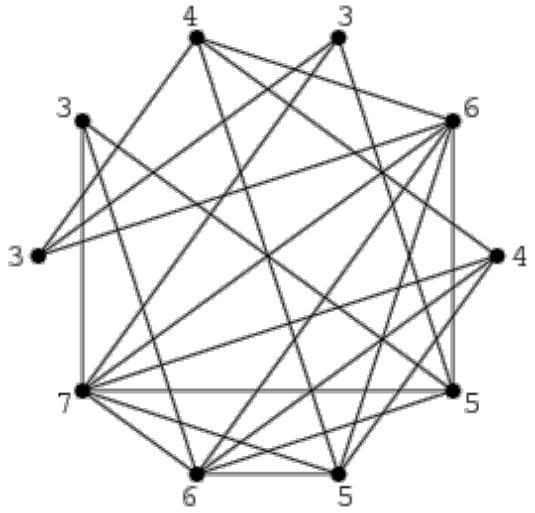
- A network (or graph) is:
  - a collection of individuals or entities, each called a vertex or node
  - a list of pairs of vertices that are neighbors, representing edges or links
- Examples:
  - vertices are mathematicians, edges represent coauthorship relationships
  - vertices are Facebook users, edges represent Facebook friendships
  - vertices are news articles, edges represent word overlap
- Networks can represent any **binary** relationship over individuals
- Often helpful to visualize networks with a diagram, but the network is the list of vertices and edges, not the visualization
  - same network has many different visualizations



Networks: Basic Definitions

- We will use the following notations:
  - n to denote the number of vertices, |V|, in a network
  - m to denote the number of edges, |E|, in a network
- Number of possible edges:
  - In an undirected network, m is at most n(n-1)/2.
- The **degree** of a vertex is its number of neighbors
  - Sum of all degrees is 2m:

$$\sum_{v \in V} \deg(v) = 2|E| \,.$$



Weisstein, Eric W. "Vertex Degree." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/VertexDegree.html

Networks: Basic Definitions

- The **distance** between two vertices is the length of the shortest path connecting them.
  - This assumes the network has only a single connected component
  - If two vertices are in different components, their distance is infinite

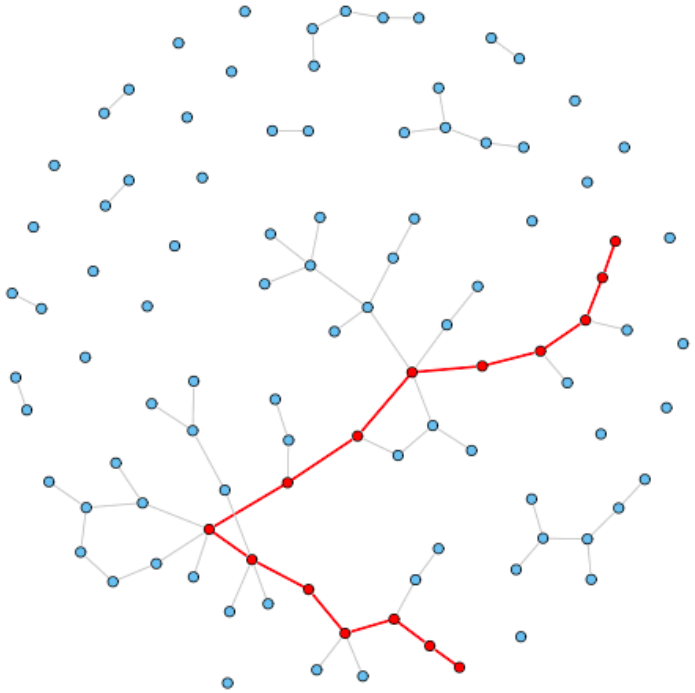
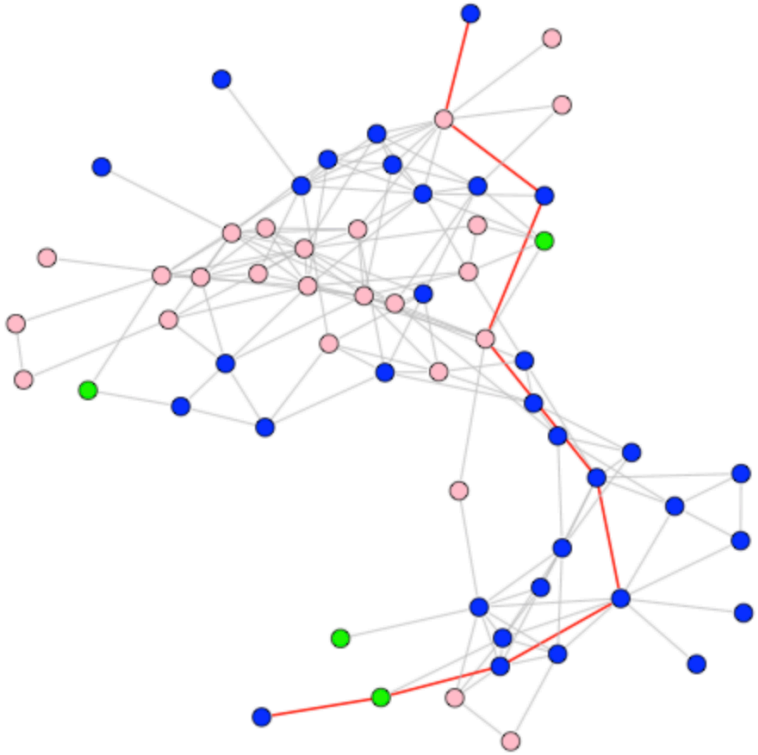


Image from
https://www.sci.unich.it/~francesc/teaching/network/geodesic.html

Networks: Basic Definitions

- The **diameter** of a network is the maximum distance between a pair of vertices in the network
  - It measures how near or far typical individuals are from each other



The dolphin network with the diameter (the longest shortest path) highlighted in red. The diameter is 8 edges long.

From https://users.dimi.uniud.it/~massimo.franceschet/bottlenose/bottlenose.html

Networks: Basic Definitions

- So far, we have been discussing undirected networks
- Connection relationship is symmetric:
  - if vertex u is connected to vertex v, then v is also connected to u
  - Facebook friendship is symmetric/reciprocal
- Sometimes we'll want to discuss **directed** networks
  - I can follow you on Twitter without you following me
  - web page A may link to page B, but not vice-versa
- In such cases, directionality matters and edges are annotated by arrows

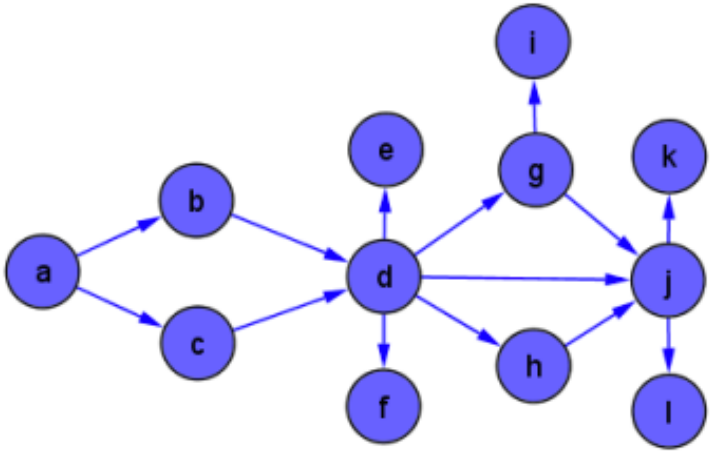
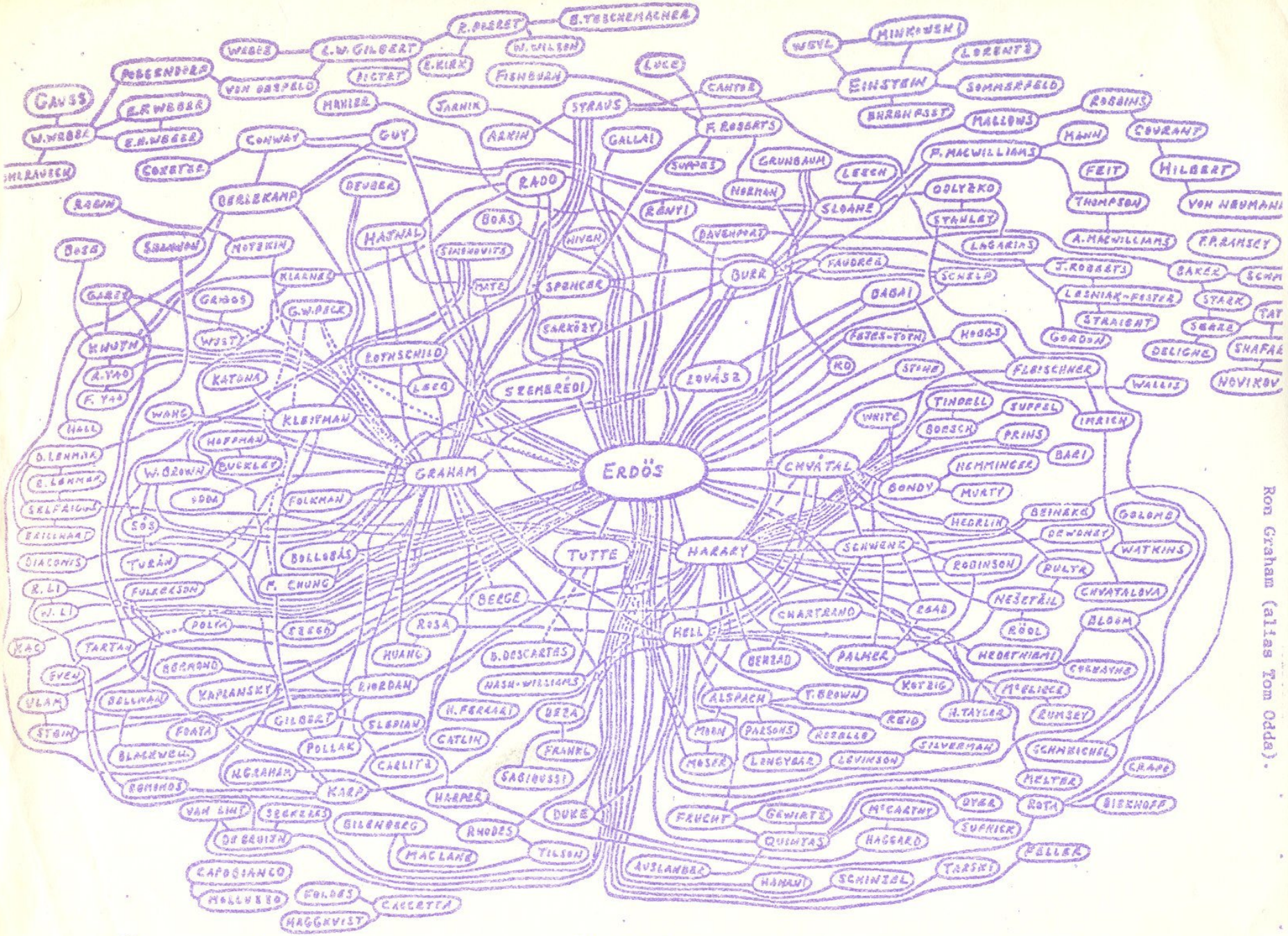


Image from https://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/index.html

Illustrating the Concepts

- Example: scientific collaboration
  - vertices: math and computer science researchers
  - links: between coauthors on a published paper
  - ***Erdős numbers***: distance to **Paul Erdős**
  - Erdős was definitely a *hub* or *connector;*
    - he had 507 coauthors
  - MG's Erdős number is 3, and there are 11 different 3-hop paths:

1. Erdős->Avis->Snoeyink->Goodrich
2. Erdős->Pach->Bronnimann->Goodrich
3. Erdős->Pollack->Agarwal->Goodrich
4. Erdős->Alon->Vishkin->Goodrich
5. Erdős->Aronov->Kosaraju->Goodrich
6. Erdős->Wagstaff->Atallah->Goodrich
7. Erdős->Silverman->Mount->Goodrich
8. Erdős->Fraenkel->Scheinerman->Goodrich
9. Erdős->Odlyzko->Guibas->Goodrich
10. Erdős->Yao->Eppstein->Goodrich (Prof. Eppstein's Erdős number is 2)
11. Erdős->Fishburn->Tanenbaum->Goodrich



Ron Graham (alias Tom Odda).

Figure 1

To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

Erdös number 0 --- 1 person
Erdös number 1 --- 504 people
Erdös number 2 --- 6593 people
Erdös number 3 --- 33605 people
Erdös number 4 --- 83642 people
Erdös number 5 --- 87760 people
Erdös number 6 --- 40014 people
Erdös number 7 --- 11591 people
Erdös number 8 --- 3146 people
Erdös number 9 --- 819 people
Erdös number 10 --- 244 people
Erdös number 11 --- 68 people
Erdös number 12 --- 23 people
Erdös number 13 --- 5 people

The median Erdös number is 5; the mean is 4.65, and the standard deviation is 1.21.

The Kevin Bacon Game

Boxed version of the Kevin Bacon Game

Invented by Albright College students in 1994:

- Craig Fass, Brian Turtle, Mike Ginelly

Goal: Connect any actor to Kevin Bacon, by linking actors who have acted in the same movie.

Oracle of Bacon website uses Internet Movie Database (IMDB.com) to find shortest link between any two actors:

http://oracleofbacon.org/

The Kevin Bacon Game

An Example

Kevin Bacon

Mystic River (2003)

Tim Robbins

Code 46 (2003)

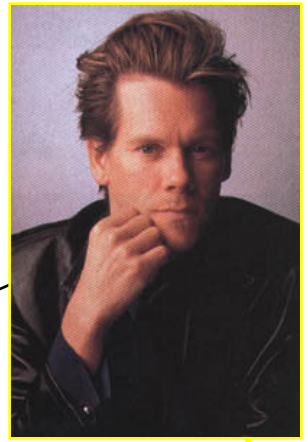
Om Puri

Yuva (2004)

Rani Mukherjee

Black (2005)

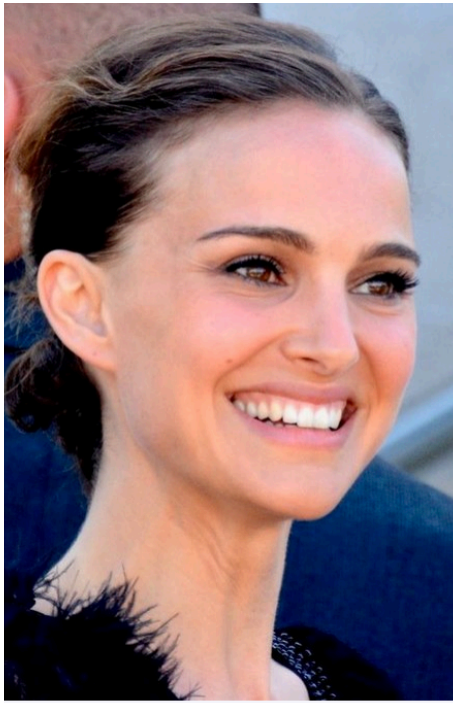
Amitabh Bachchan



Erdős-Bacon Numbers

- The sum of a person's Erdős number and their Bacon number.

Colin Firth's is 6=5+1

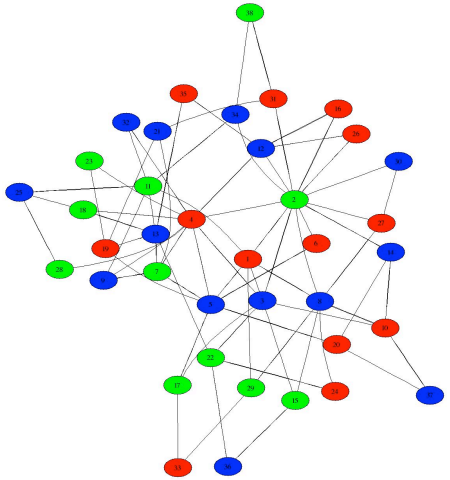


Natalie Portman's is 7=5+2

Network *Structure*

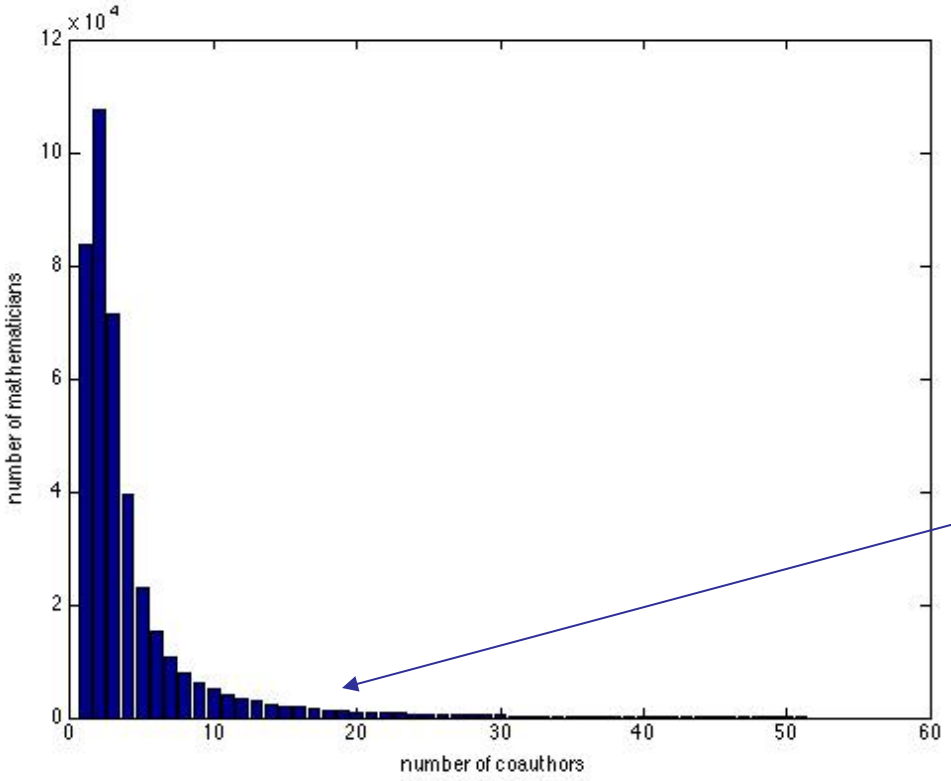
- Emphasize purely *structural* properties
  - size, diameter, connectivity, degree distribution, etc.
  - may examine statistics across many networks
  - will also use the term *topology* to refer to structure
- Structure can reveal:
  - community
  - “important” vertices, centrality, etc.
  - robustness and vulnerabilities
  - can also impose *constraints* on dynamics
- Less emphasis on what actually occurs *on* network
  - web pages are linked… but people surf the web
  - buyers and sellers exchange goods and cash
  - friends are connected… but have specific interactions

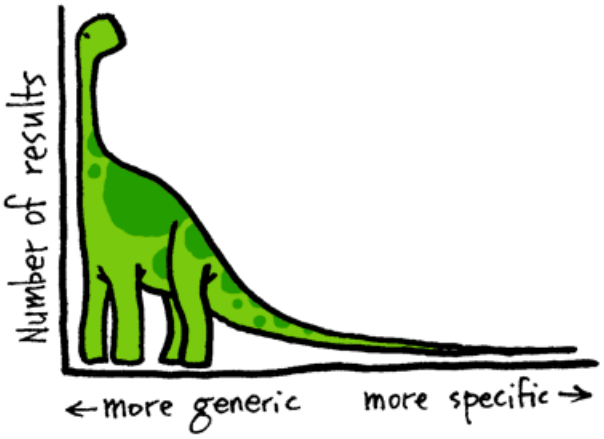
Network Structure:
1. Power Law Degree Distributions



Math Collaboration Degree Distribution

- x axis: number of neighbors/coauthors (degree)
- y axis: number of mathematicians with that degree





A long tail
(also known as a heavy tail)

What Do We Mean By *Not* "Heavy-Tailed"?

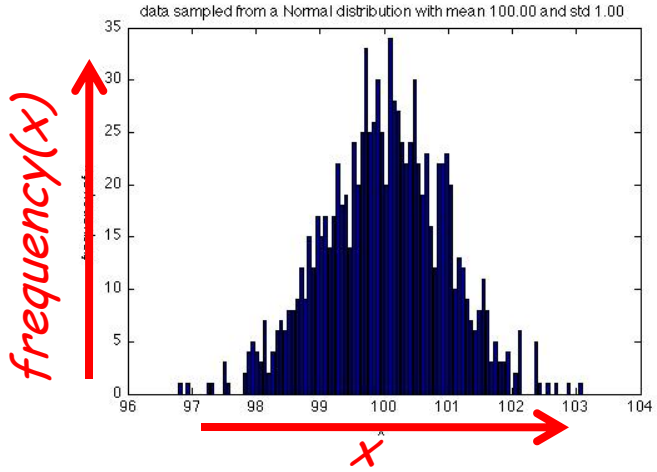
- Mathematical model of a typical "bell-shaped" distribution:
  - the Normal or Gaussian distribution over some quantity x
  - Good for modeling many real-world quantities… but not degree distributions
  - if mean/average is μ then probability of value x is:

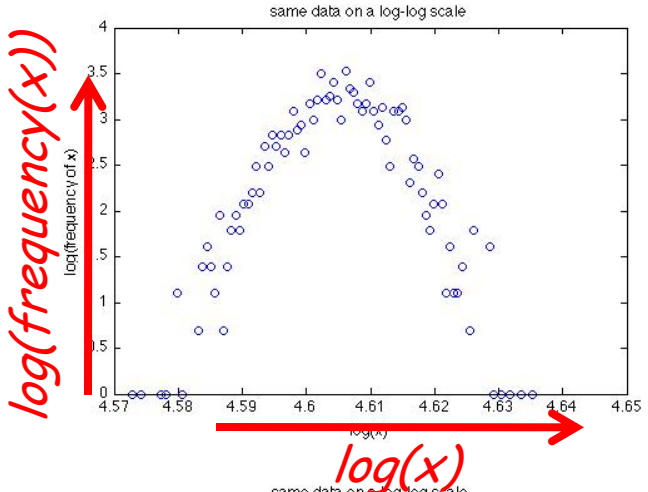
$$probability(x) \propto e^{-(x-\mu)^2}$$

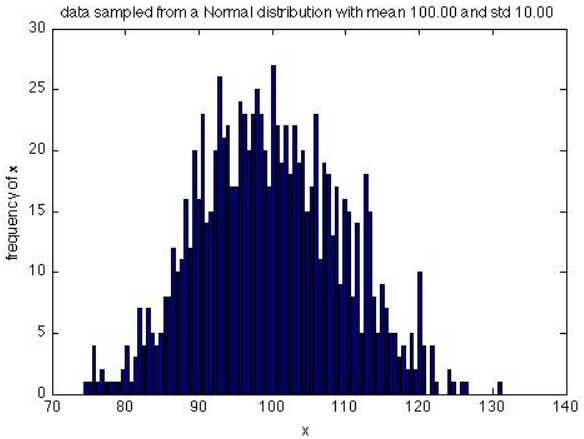
  - main point: exponentially fast decay as x moves away from μ
  - if we take the logarithm:

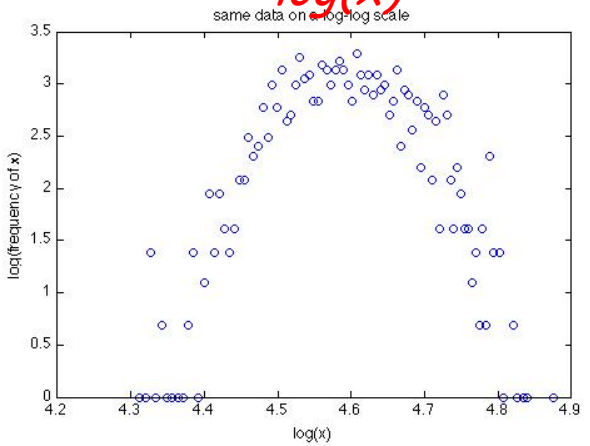
$$\log(probability(x)) \propto -(x-\mu)^2$$

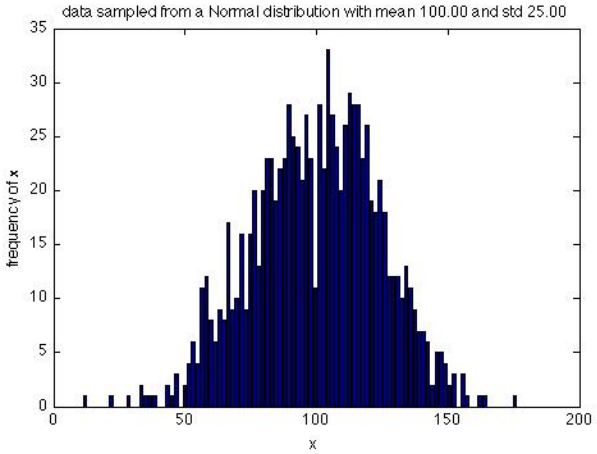
- Claim: if we plot log(x) vs log(probability(x)), will get strong curvature
- Let's look at some (artificial) sample data…
  - (Poisson better than Normal for degrees, but same story holds)

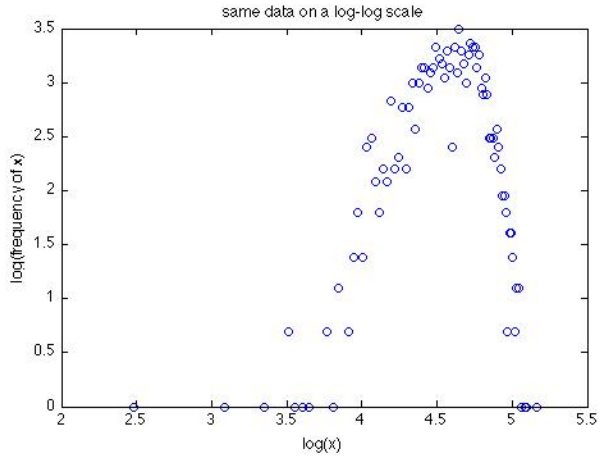

data sampled from a Normal distribution with mean 100.00 and std 1.00
frequency(x)
x


same data on a log-log scale
log(frequency(x))
log(x)


data sampled from a Normal distribution with mean 100.00 and std 10.00
frequency of x
x


same data on a log-log scale
log(frequency of x)
log(x)


data sampled from a Normal distribution with mean 100.00 and std 25.00
frequency of x
x


same data on a log-log scale
log(frequency of x)
log(x)

What *Do* We Mean By "Heavy-Tailed"?

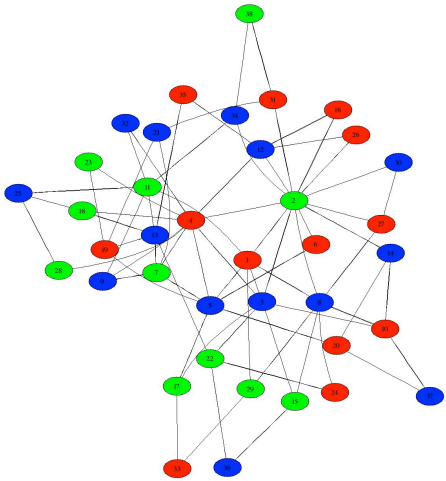
- One mathematical model of a typical "heavy-tailed" distribution:
  - the Power Law distribution with exponent β

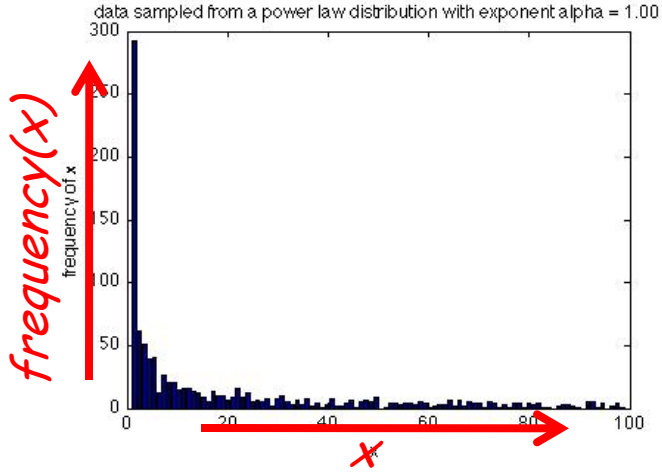
$$probability(x) \propto 1/x^{\beta}$$

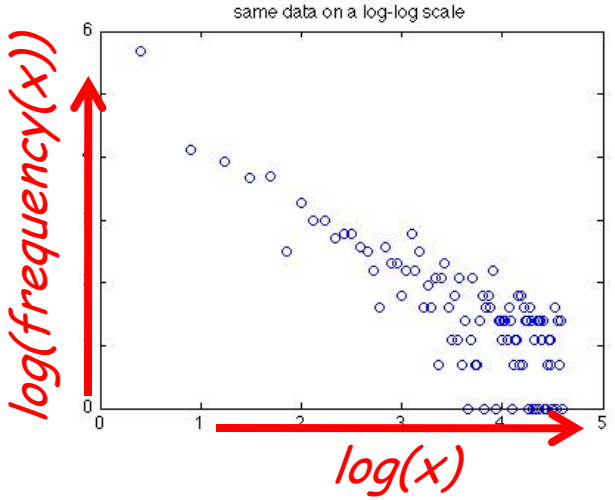
  - main point: inverse polynomial decay as x increases
  - if we take the logarithm:

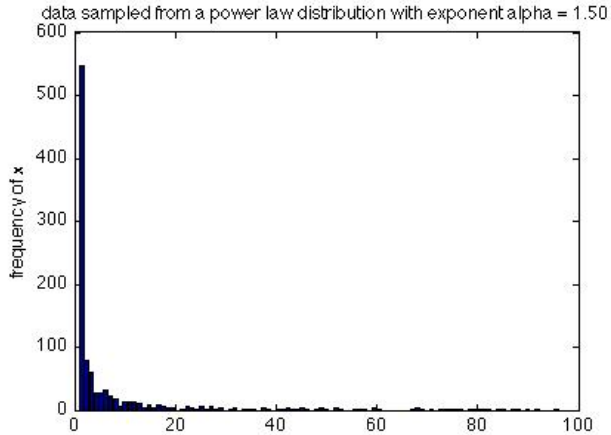
$$\log(probability(x)) \propto -\beta \log(x)$$

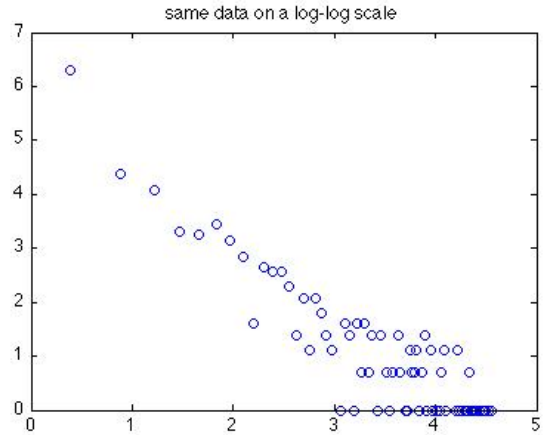
- Claim: if we plot log(x) vs log(probability(x)), will get a straight line!
- Let's look at (artificial) some sample data…

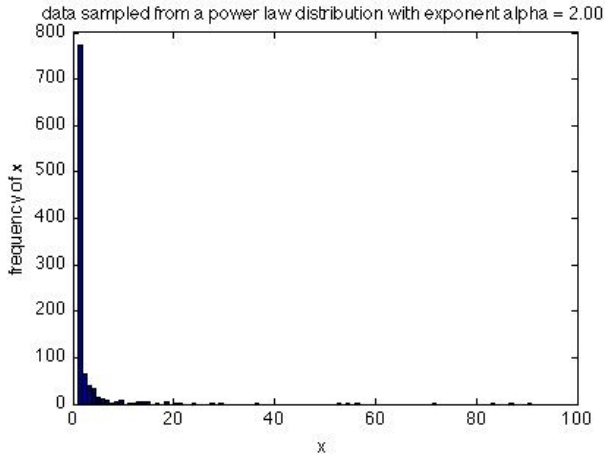


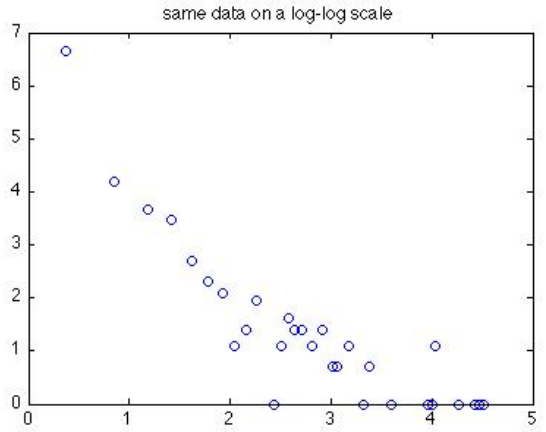

data sampled from a power law distribution with exponent alpha = 1.00
frequency of x
x
frequency(x)

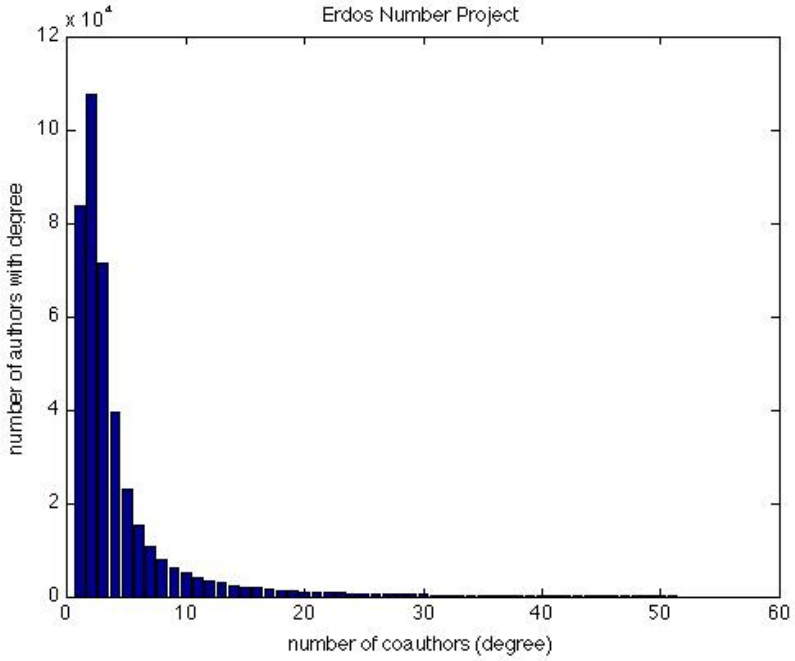

same data on a log-log scale
log(frequency(x))
log(x)

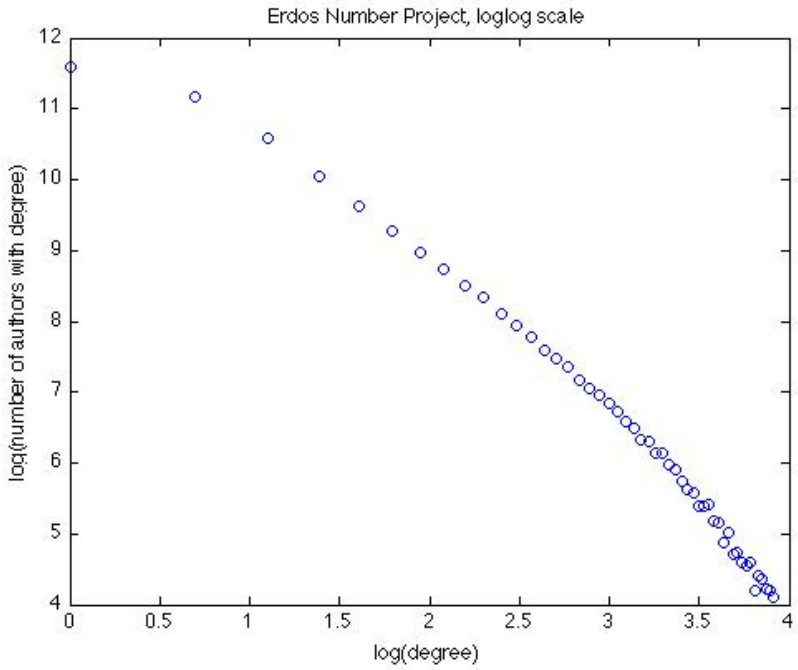

data sampled from a power law distribution with exponent alpha = 1.50
frequency of x


same data on a log-log scale

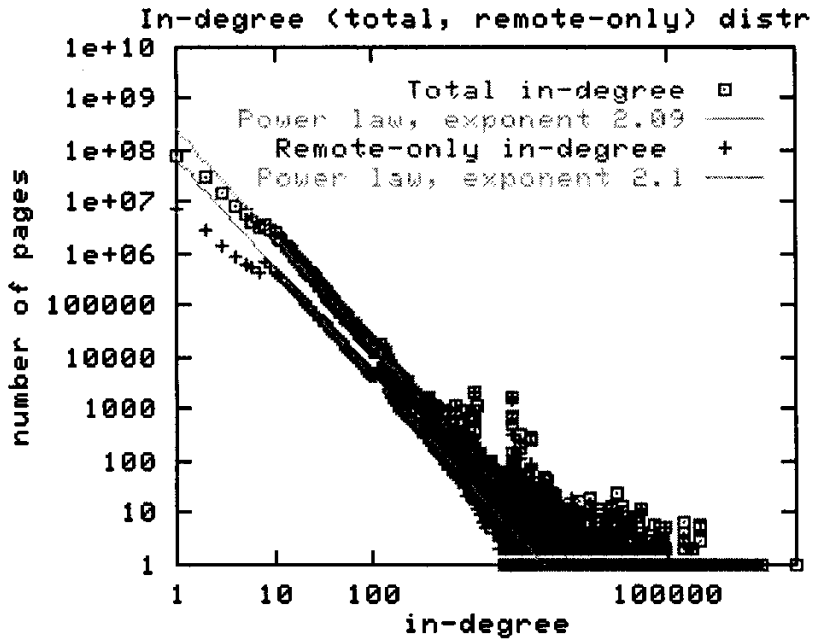

data sampled from a power law distribution with exponent alpha = 2.00
frequency of x
x

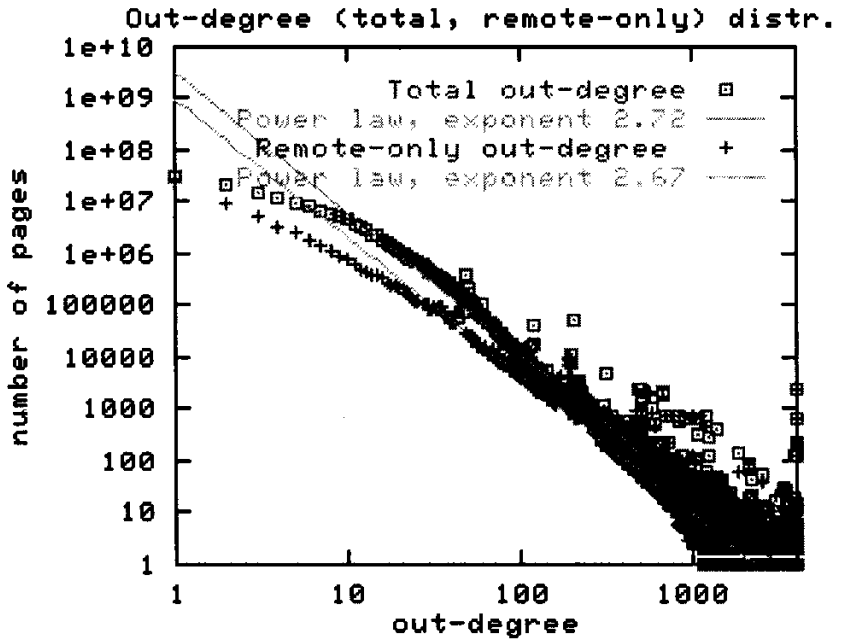

same data on a log-log scale





Erdos Number Project Revisited





Figures 1 and 2: In-degree and out-degree distributions subscribe to the power law. The law also holds if only off-site (or "remote-only") edges are considered.

Degree Distribution of the Web Graph [Broder et al.]

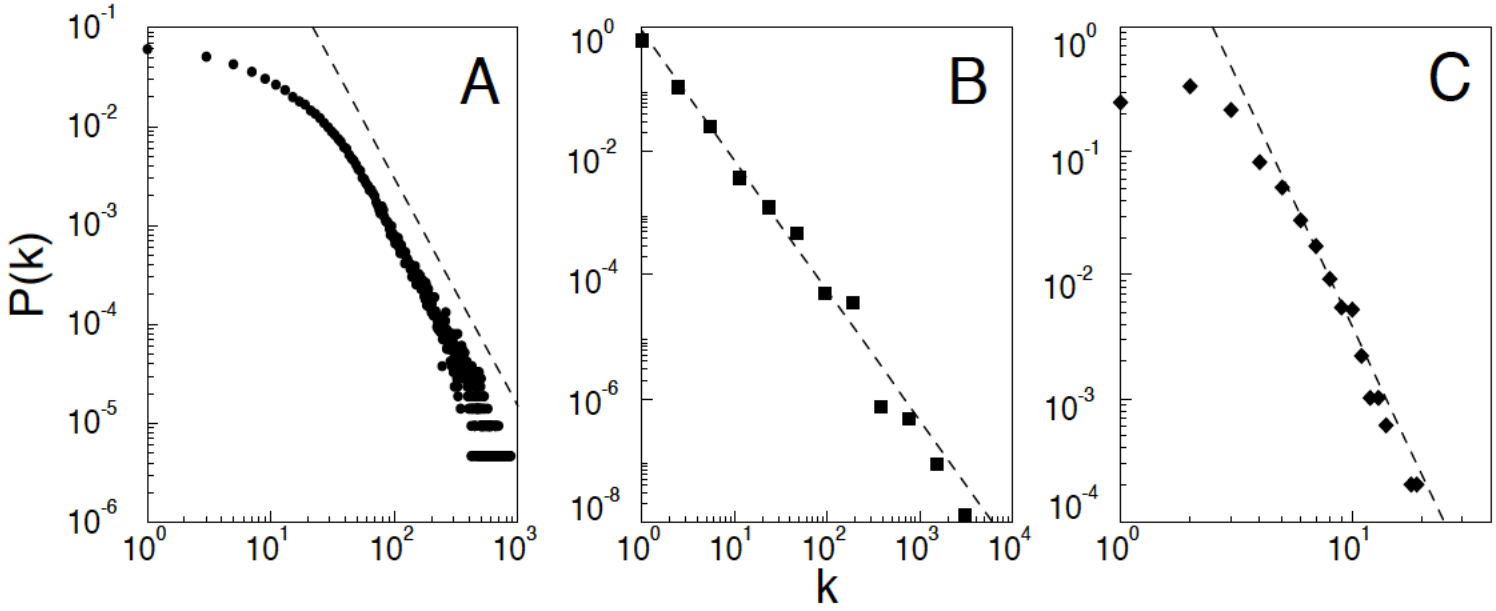
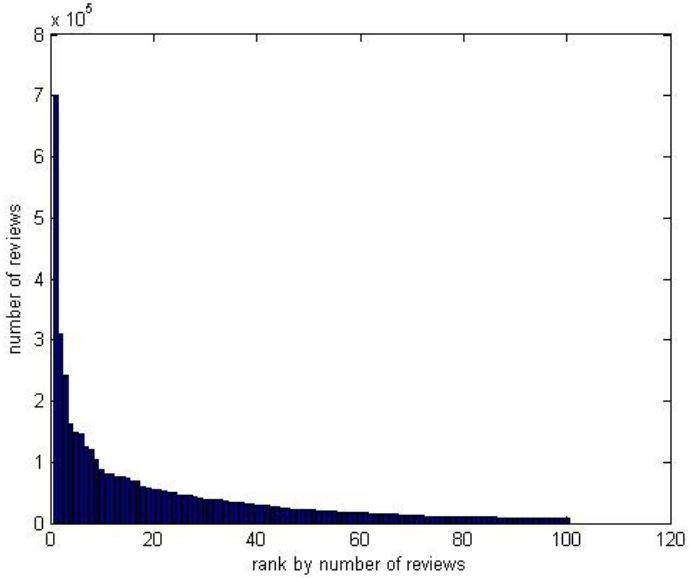


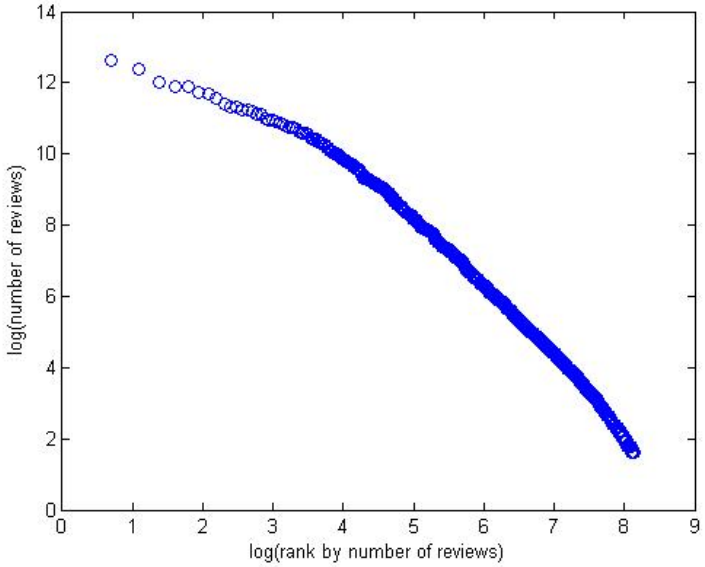
FIG. 1. The distribution function of connectivities for various large networks. (**A**) Actor collaboration graph with $N = 212,250$ vertices and average connectivity $\langle k \rangle = 28.78$; (**B**) World wide web, $N = 325,729$, $\langle k \rangle = 5.46$ (*6*); (**C**) Powergrid data, $N = 4,941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (**A**) $\gamma_{actor} = 2.3$, (**B**) $\gamma_{www} = 2.1$ and (**C**) $\gamma_{power} = 4$.

Actor Collaborations; Web; Power Grid [Barabasi and Albert]

Power Laws

- A distribution has a **power law** if
  - $P(k) = ck^{-\alpha}$, for constants, c and α
  - So $\log P(k) = \log c + -\alpha \log k$
  - Substituting $b = \log c$, $y = \log P(k)$, and $x = \log k$,
  - $y = -\alpha x + b$
- Look at the frequency of English words:
  - "the" is the most common, followed by "of", "to", etc.
  - claim: frequency of the n-th most common ~ 1/n (power law, a ~ 1)
- General theme:
  - *rank* events by their *frequency of occurrence*
  - resulting distribution often is a power law!
- Other examples:
  - North America city sizes
  - personal income
  - file sizes
  - let's look at log-log plots of these

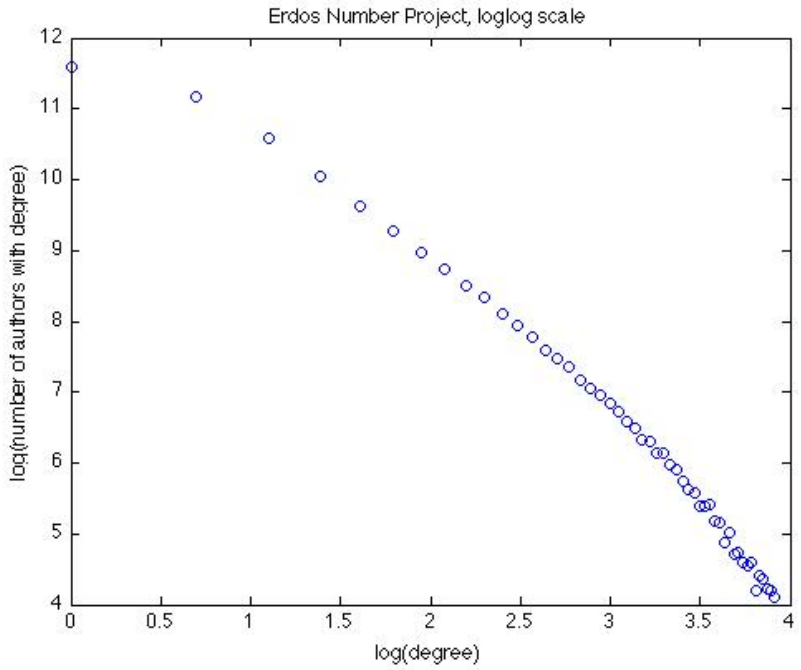




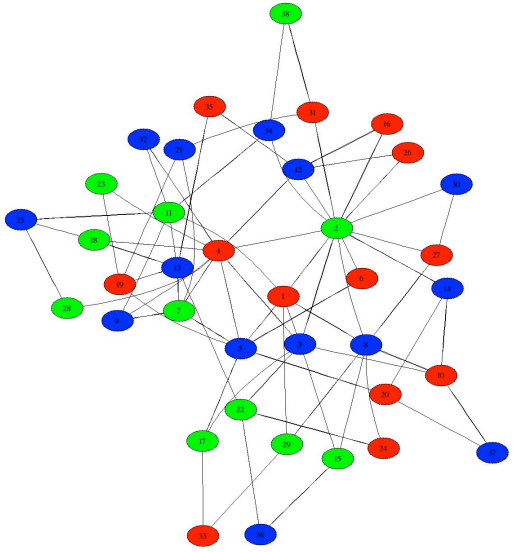
iPhone App Popularity

Power Laws

- Power law distribution is a good mathematical model for heavy tails; Normal/bell-shaped is not
- Statistical signature of power law and heavy tails: linear on a log-log scale
- Many social and other networks exhibit this signature

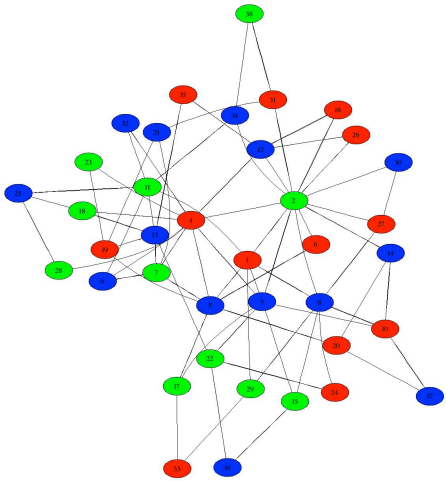


Network Structure:
2. Small Diameter



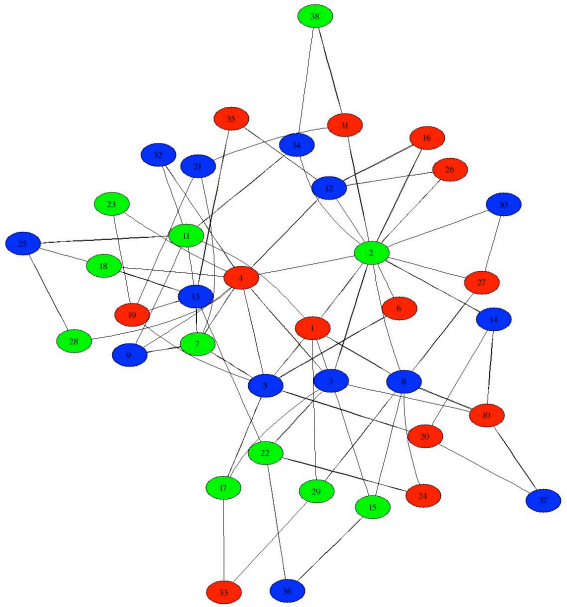
What Do We Mean By "Small Diameter"?

- First let's recall the definition of diameter:
  - assumes network has a single connected component (or examine "giant" component)
  - for every pair of vertices u and v, compute shortest-path distance d(u,v)
  - equivalent: pick a random pair of vertices (u,v); what do we expect d(u,v) to be?
- What's the smallest/largest diameter(G) could be?
  - smallest: 1 (complete network, all N(N-1)/2 edges present); independent of N
  - largest: linear in N (chain or line network)
- "Small" diameter:
  - no precise definition, but certainly << N
  - Travers and Milgram: ~5; any fixed network has fixed diameter
  - may want to allow diameter to grow slowly with N (?)
  - e.g. log(N)

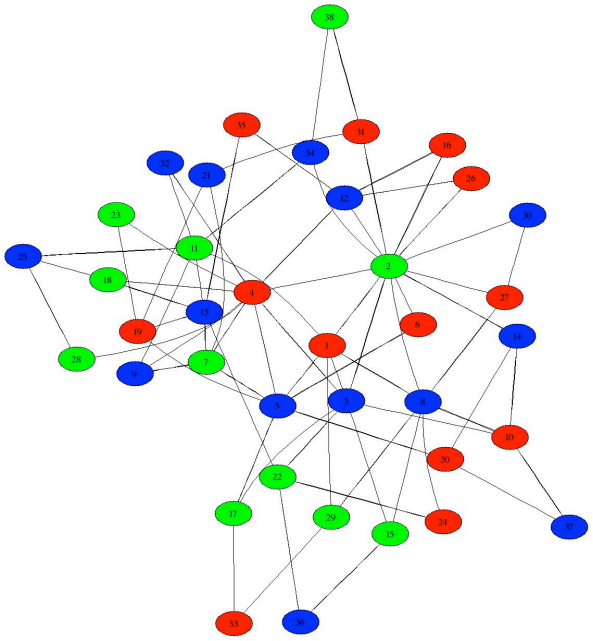


Empirical Support for Small Diameter

- Travers and Milgram, 1969:
  - diameter ~ 5-6, N ~ 200M
- Columbia Small Worlds, 2003:
  - diameter ~4-7, N ~ web population?
- Lescovec and Horvitz, 2008:
  - Microsoft Messenger network
  - Diameter ~6.5, N ~ 180M
- Backstrom et al., 2012:
  - Facebook social graph
  - diameter ~5, N ~ 721M



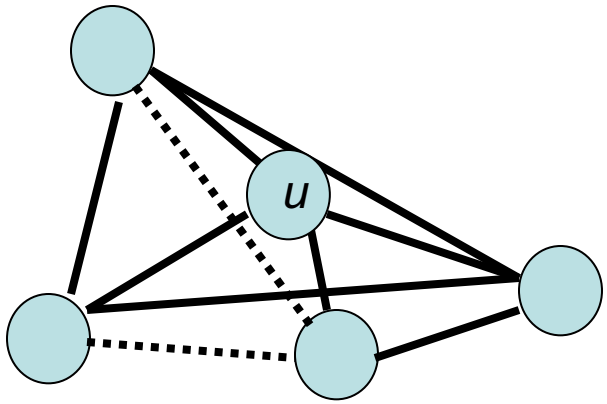
Network Structure:
3. Clustering of Connectivity



The Clustering Coefficient of a Network

- Intuition: a measure of how "bunched up" edges are
- The clustering coefficient of vertex u:
  - let k = degree of u = number of neighbors of u
  - k(k-1)/2 = *max possible # of edges* between neighbors of u
  - c(u) = (*actual* # of edges between neighbors of u)/[k(k-1)/2]
  - fraction of pairs of friends that are also friends
  - 0 <= c(u) <= 1; measure of *cliquishness* of u's neighborhood
- Clustering coefficient of a graph G:
  - CC(G) = average of c(u) over all vertices u in G

k = 4
k(k-1)/2 = 6
c(u) = 4/6 = 0.666…



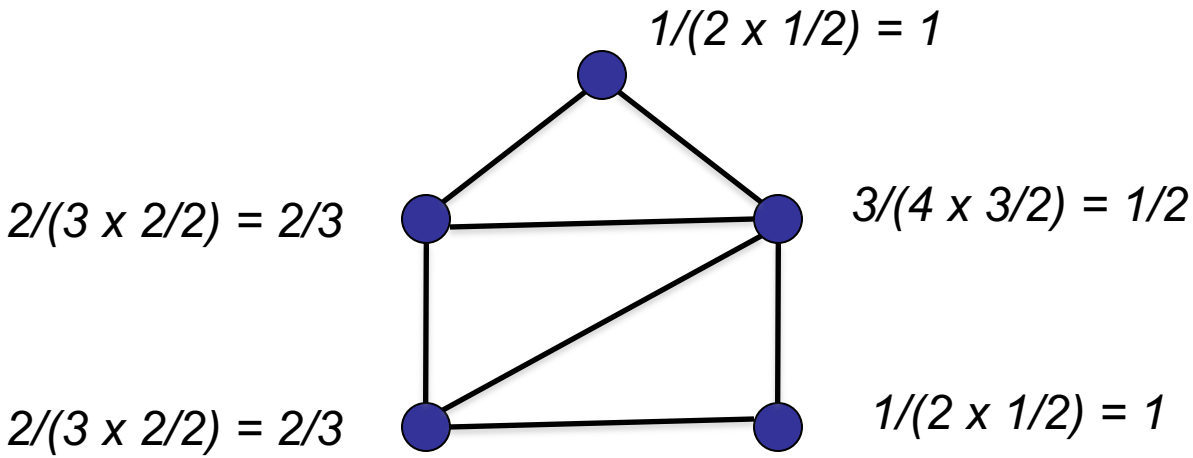
What Do We Mean By "High" Clustering?

- CC(G) measures how likely vertices with a common neighbor are to be neighbors themselves
- Should be compared to how likely *random* pairs of vertices are to be neighbors
- Let p be the edge density of network/graph G:

$$p = E / (N(N-1)/2)$$

- Here E = total number of edges in G
- If we picked a pair of vertices at random in G, probability they are connected is exactly p
- So we will say clustering is high if CC(G) >> p

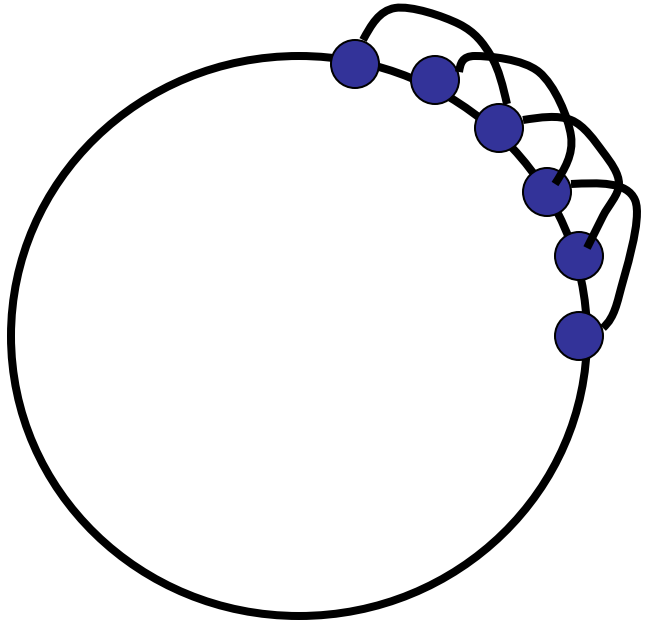
Clustering Coefficient Example 1



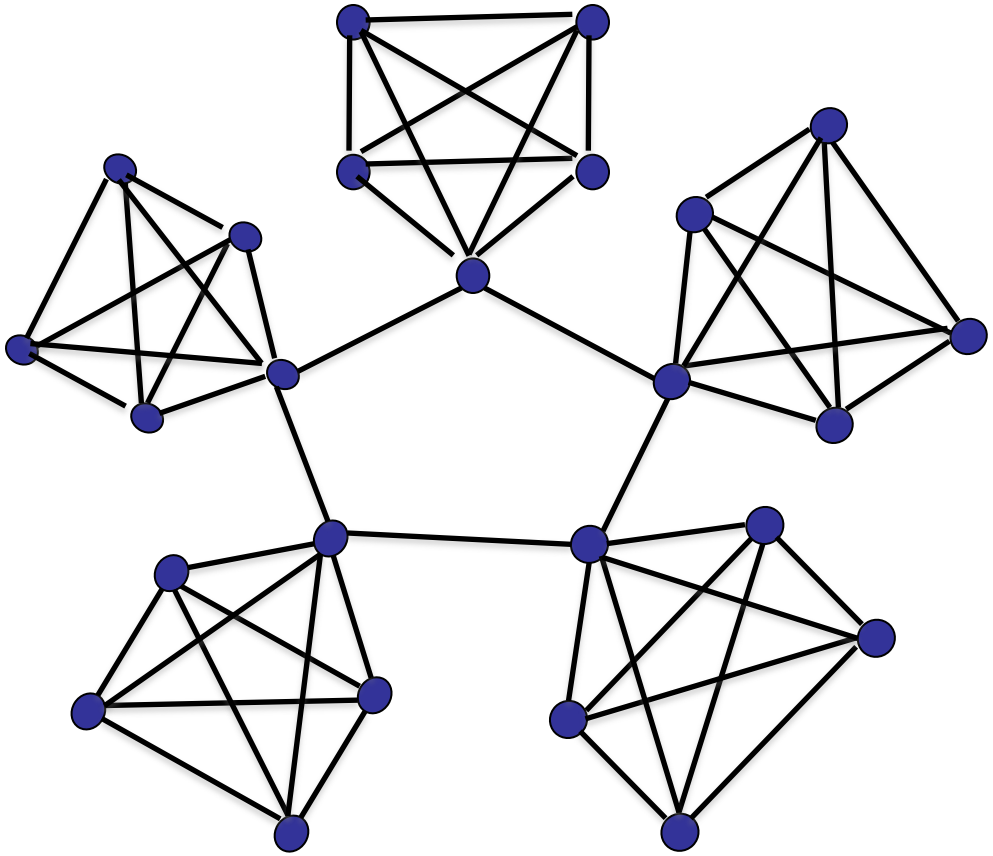
C.C. = (1 + ½ + 1 + 2/3 + 2/3)/5 = 0.7666…
p = 7/(5 x 4/2) = 0.7
Not highly clustered

Clustering Coefficient Example 2

- Network: simple cycle + edges to vertices 2 hops away on cycle
- By symmetry, all vertices have the same clustering coefficient
- Clustering coefficient of a vertex v:
  - Degree of v is 4, so the number of *possible* edges between pairs of neighbors of v is 4 x 3/2 = 6
  - How many pairs of v's neighbors actually *are* connected? 3 --- the two clockwise neighbors, the two counterclockwise, and the immediate cycle neighbors
  - So the c.c. of v is 3/6 = ½
- Compare to overall edge density:
  - Total number of edges = 2N
  - Edge density p = 2N/(N(N-1)/2) ~ 4/N
  - As N becomes large, ½ >> 4/N
  - So this cyclical network is highly clustered



Clustering Coefficient Example 3



Divide N vertices into sqrt(N) groups of size sqrt(N) (here N = 25)
Add all connections within each group (cliques), connect "leaders" in a cycle
N – sqrt(N) non-leaders have C.C. = 1, so network C.C. → 1 as N becomes large
Edge density is p ~ 1/sqrt(N)