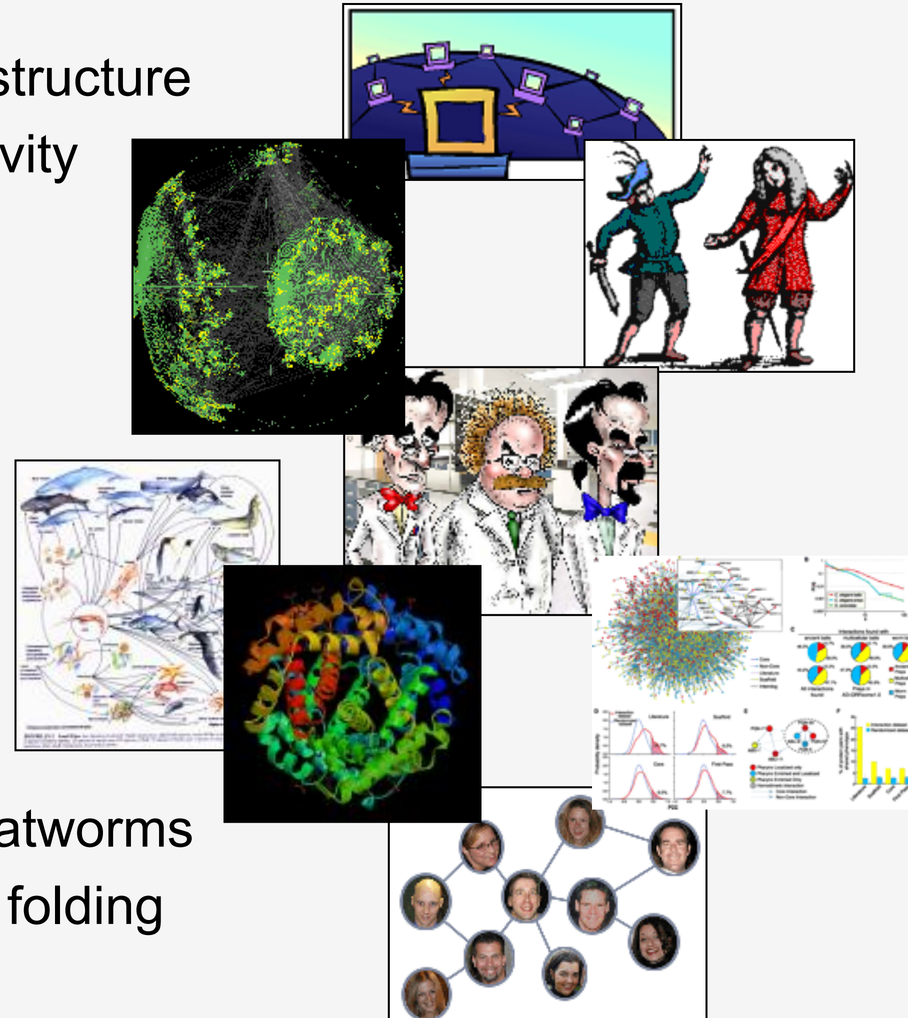
Social Network Analysis

Algorithms and Applications

Arindam Pal
TCS Research & Innovation
arindamp@gmail.com

APPLICATIONS OF NETWORK THEORY

- World Wide Web and hyperlink structure
- The Internet and router connectivity
- Collaborations among
  - Movie actors
  - Scientists and mathematicians
- Romantic relationships
- Cellular networks in biology
- Food webs in ecology
- Phone call patterns
- Word co-occurrence in text
- Neural network connectivity of flatworms
- Conformational states in protein folding

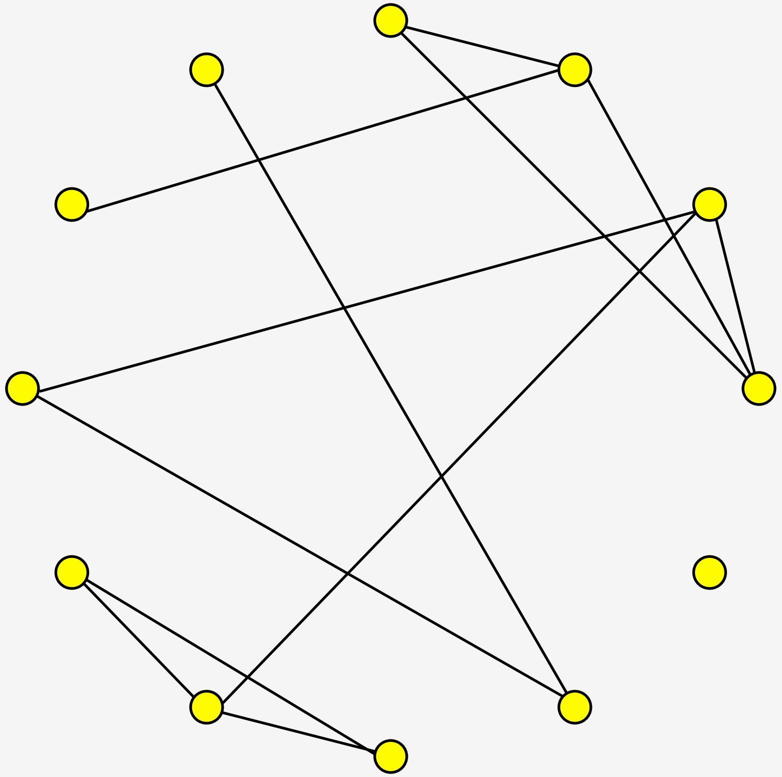


WEB APPLICATIONS OF SOCIAL NETWORKS

- Analyzing page importance
  - Page Rank
    - Related to recursive in-degree computation
  - Authorities/Hubs
- Discovering Communities
  - Finding near-cliques
- Analyzing Trust
  - Propagating Trust
  - Using propagated trust to fight spam
    - In Email
    - In Web page ranking

SOCIETY AS A GRAPH

- People are represented as *nodes.*
- Relationships are represented as *edges.*
  - (Relationships may be acquaintanceship, friendship, co-authorship, etc.)
- Allows analysis using tools of graph theory



HISTORY

- 17th century: Spinoza developed first model
- 1937: J.L. Moreno introduced sociometry; he also invented the sociogram
- 1948: A. Bavelas founded the group networks laboratory at MIT; he also specified centrality

HISTORY

- 1949: A. Rapaport developed a probability based model of information flow
- 50s and 60s: Distinct research by individual researchers
- 70s: Field of social network analysis emerged.
  - New features in graph theory – more general structural models
  - Better computational power – analysis of complex relational data sets

CONNECTIONS

- Size
  - Number of nodes
- Density
  - Number of ties that are present the amount of ties that could be present
- Out-degree
  - Sum of connections from an actor to others
- In-degree
  - Sum of connections to an actor

DISTANCE

- Walk
  - A sequence of actors and relations that begins and ends with actors
- Geodesic distance
  - The number of relations in the shortest possible walk from one actor to another
- Maximum flow
  - The amount of different actors in the neighborhood of a source that lead to pathways to a target

SOME MEASURES OF POWER & PRESTIGE

- Degree
  - Sum of connections from or to an actor
    - Transitive weighted degree→Authority, hub, pagerank
- Closeness centrality
  - Distance of one actor to all others in the network
- Betweenness centrality
  - Number that represents how frequently an actor is between other actors' geodesic paths

CLIQUES AND SOCIAL ROLES

- Cliques
  - Sub-set of actors
    - More closely tied to each other than to actors who are not part of the sub-set
      - (A lot of work on “trawling” for communities in the web-graph)
      - Often, you first find the clique (or a densely connected subgraph) and then try to interpret what the clique is about

- Social roles
  - Defined by regularities in the patterns of relations among actors

OUTLINE

Small Worlds

Random Graphs

Alpha and Beta

Power Laws

Searchable Networks

Six Degrees of Separation

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)

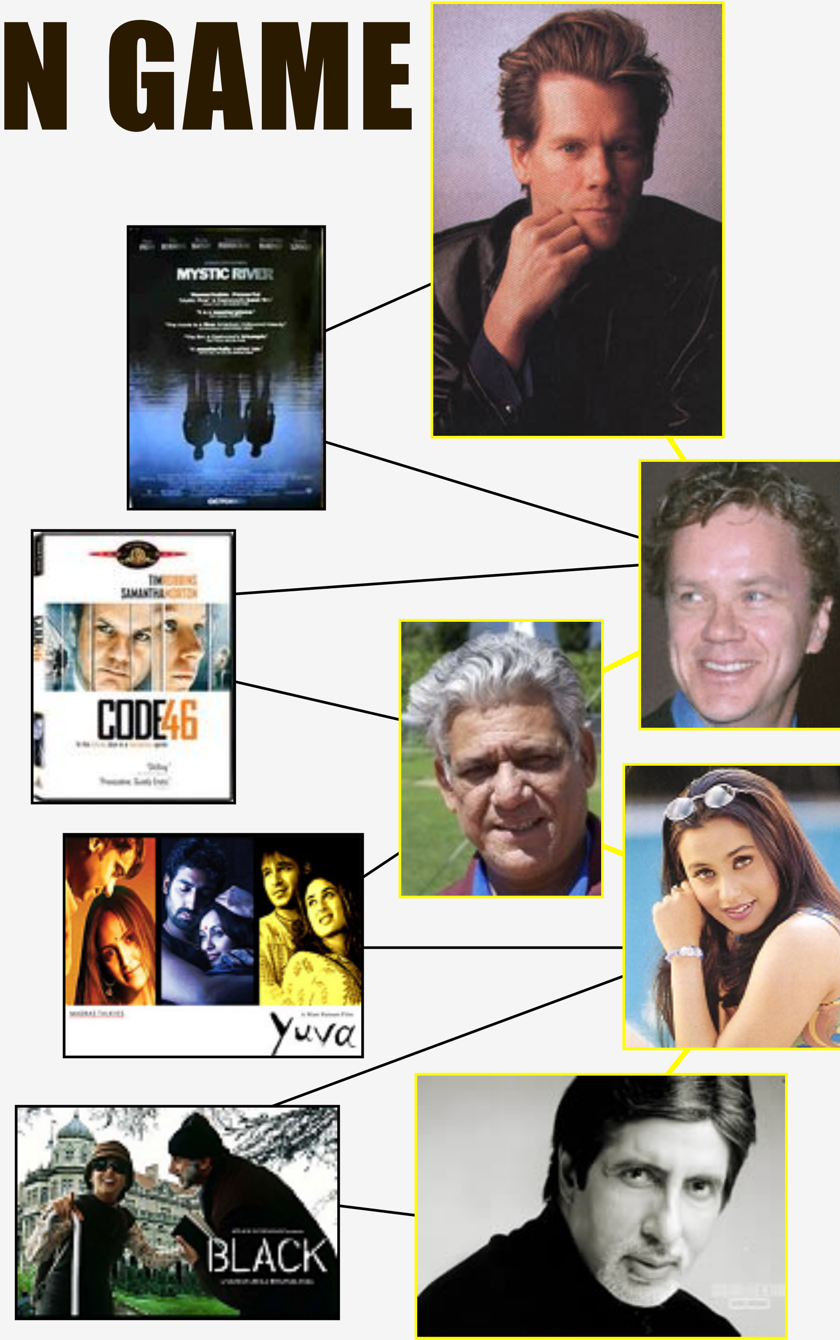
Om Puri

Yuva (2004)

Rani Mukherjee

Black (2005)

Amitabh Bachchan



ACTUALLY AMITABH BACHCHAN HAS A BACON NUMBER 2

THE KEVIN BACON GAME

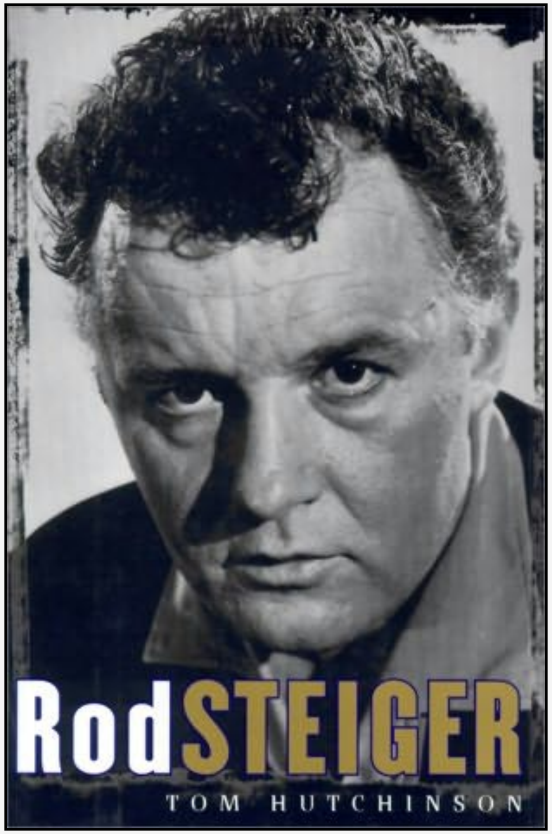
Total # of actors in database: *~550,000*

Average path length to Kevin: *2.79*

Actor closest to “center”: *Rod Steiger (2.53)*

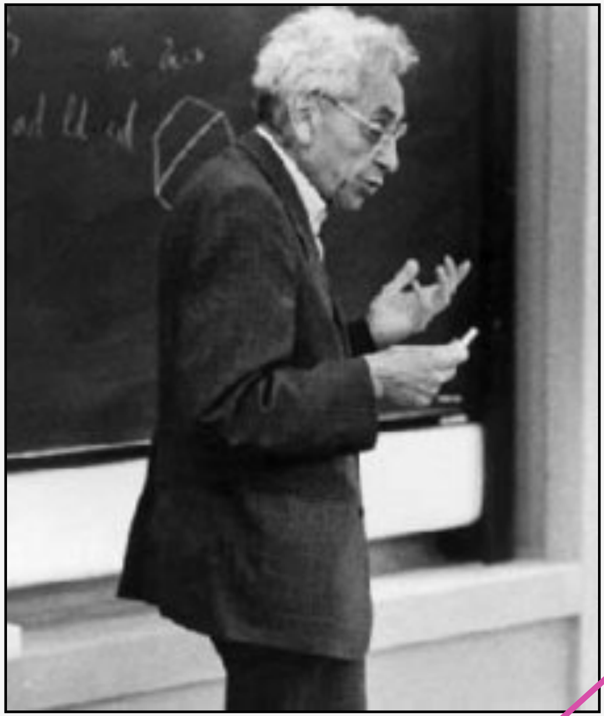
Rank of Kevin, in closeness to center: *876th*

Most actors are within three links of each other!



Center of Hollywood?

ERDŐS NUMBER (BACON GAME FOR RESEARCHERS ☺)



Paul Erdős (1913-1996)

Number of links required to connect scholars to Erdős, via co-authorship of papers

Erdős wrote 1500+ papers with 507 co-authors.

Jerry Grossman's (Oakland Univ.) website allows mathematicians to compute their Erdos numbers:

http://www.oakland.edu/enp/

Connecting path lengths, among mathematicians only:

- average is *4.65*
- maximum is *13*

Unlike Bacon, Erdos has better centrality in his network

ERDŐS NUMBER

My Erdős number is 3.

Paul Erdős – S. B. Rao – Sushmita Ruj – Arindam Pal

SIX DEGREES OF SEPARATION: MILGRAM (1967)

The experiment:

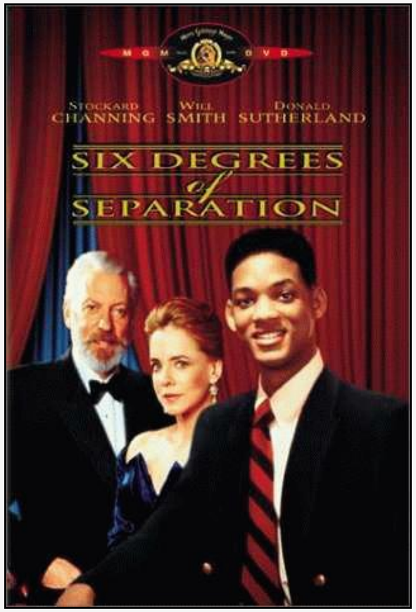
- Random people from Nebraska were to send a letter (via intermediaries) to a stock broker in Boston.

- Could only send to someone with whom they were on a first-name basis.

Among the letters that found the target, the average number of links was six.

Stanley Milgram (1933-1984)

SIX DEGREES OF SEPARATION



John Guare wrote a play called *Six Degrees of Separation*, based on this concept.

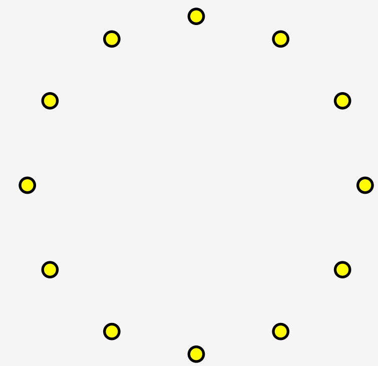
"Everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice… It's not just the big names. It's anyone. A native in a rain forest. A Tierra del Fuegan. An Eskimo. I am bound to everyone on this planet by a trail of six people…"

RANDOM GRAPHS

Erdős and Renyi (1959)

N = 12

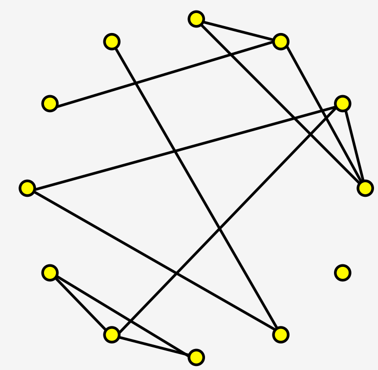
$p = 0.0$; $k = 0$



N nodes

A pair of nodes has probability p of being connected.

$p = 0.09$; $k = 1$

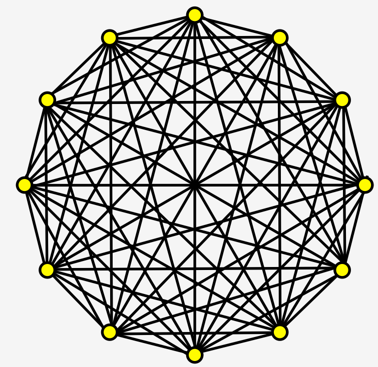


Average degree, $k = p(N - 1)$

What interesting things can be said for different values of p or k ?

(in the limit as $N \rightarrow \infty$)

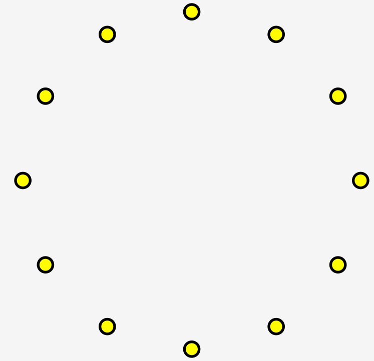
$p = 1.0$; $k \approx N$

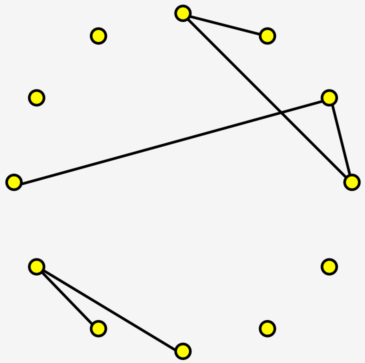


RANDOM GRAPHS

Erdős and Renyi (1959)

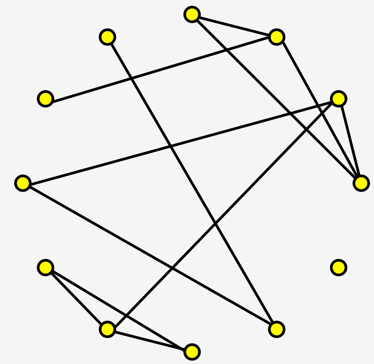
$p = 0.0$; $k = 0$





$p = 0.045$; $k = 0.5$

$p = 0.09$; $k = 1$



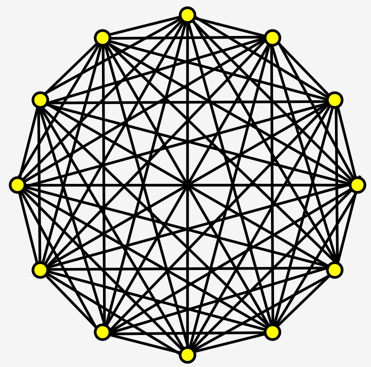
Let's look at…

Size of the largest connected cluster

Diameter (maximum path length) of the largest cluster

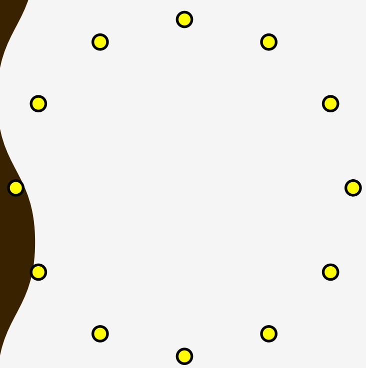
Average path length between nodes (if a path exists)

$p = 1.0$; $k = N - 1$

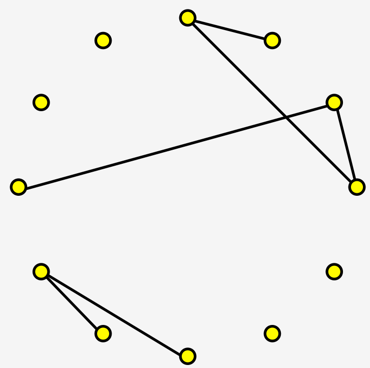


RANDOM GRAPHS

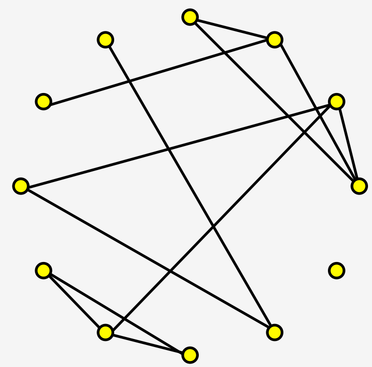
Erdős and Renyi (1959)



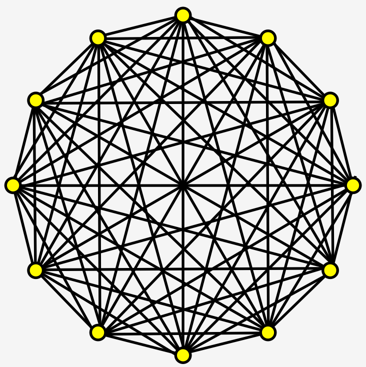
$p = 0.0$; $k = 0$



$p = 0.045$; $k = 0.5$



$p = 0.09$; $k = 1$



$p = 1.0$; $k \approx N$

Size of largest component

1	5	11	12

Diameter of largest component

0	4	7	1

Average path length between (connected) nodes

0.0	2.0	4.2	1.0

RANDOM GRAPHS

Erdős and Renyi (1959)

If $k < 1$:

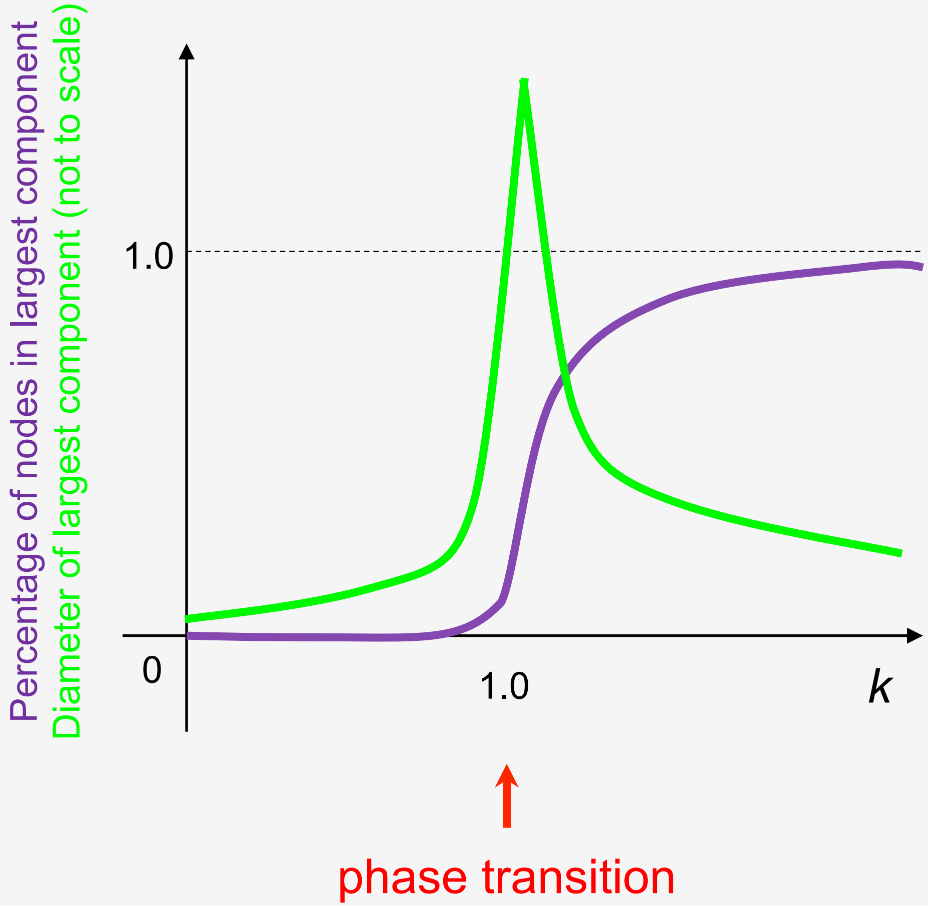
- small, isolated clusters
- small diameters
- short path lengths

At $k = 1$:

- a *giant component* appears
- diameter peaks
- path lengths are high

For $k > 1$:

- almost all nodes are connected
- diameter shrinks
- path lengths shorten



RANDOM GRAPHS

Erdős and Renyi (1959)

What does this mean?

- If connections between people can be modeled as a random graph, then…
  - Because the average person easily knows more than one person (k >> 1),
  - We live in a "small world" where within a few links, we are connected to anyone in the world.
  - Erdős and Renyi showed that average path length between connected nodes is $\frac{\ln N}{\ln k}$

RANDOM GRAPHS

Erdős and Renyi (1959)

What does this mean?

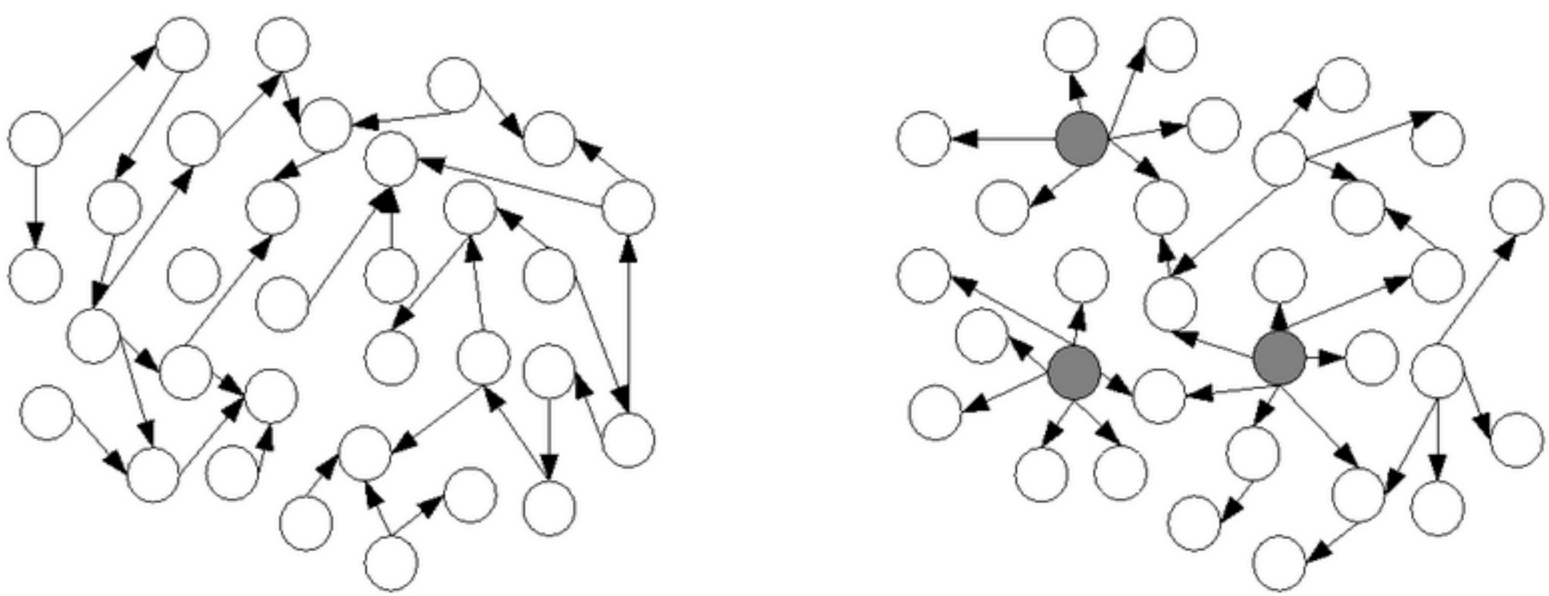
BIG "IF"!!!

- If connections between people can be modeled as a random graph, then…
  - Because the average person easily knows more than one person (k >> 1),
  - We live in a "small world" where within a few links, we are connected to anyone in the world.
  - Erdős and Renyi computed average path length between connected nodes to be: $$\frac{\ln N}{\ln k}$$

RANDOM VS. REAL SOCIAL NETWORKS

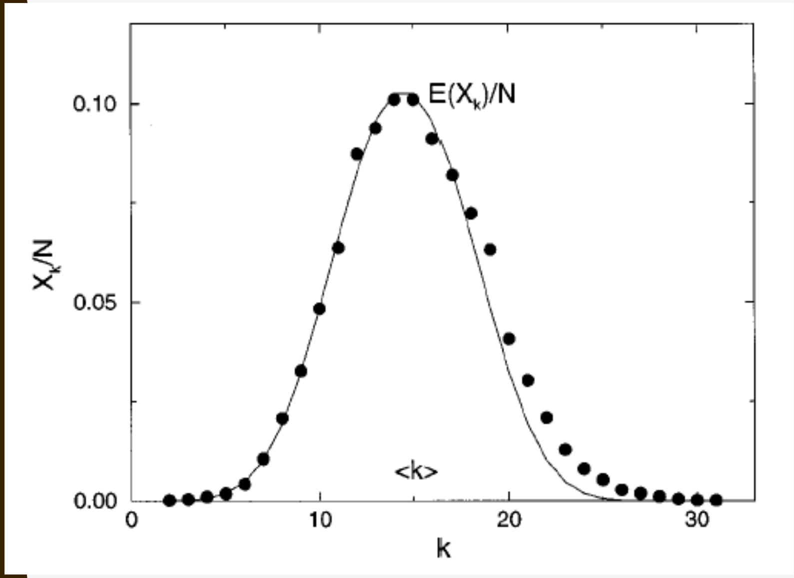
- Random network models introduce an edge between any pair of vertices with a probability *p*
  - The problem here is NOT randomness, but rather the <u>distribution</u> used (which, in this case, is *uniform*)

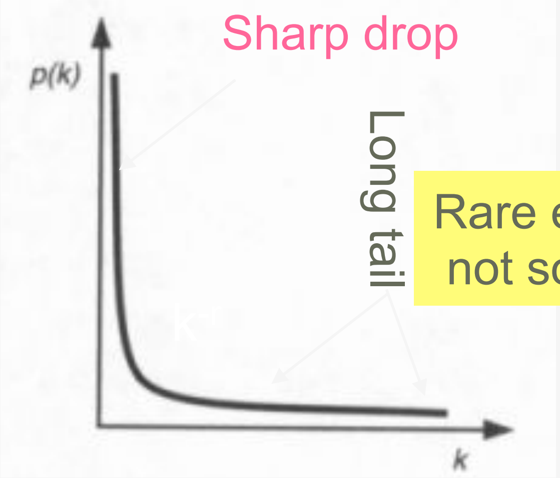
- Real networks are not exactly like these
  - Tend to have a relatively few nodes of high connectivity (the "Hub" nodes)
  - These networks are called "Scale-free" networks
    - Macro properties scale-invariant



(a) Random network (b) Scale-free network

DEGREE DISTRIBUTION & POWER LAWS





Rare events are not so rare!

Degree distribution of a random graph,
N = 10,000 p = 0.0015 k = 15.
(Curve is a Poisson curve, for comparison.)

But, many real-world networks exhibit a *power-law* distribution.

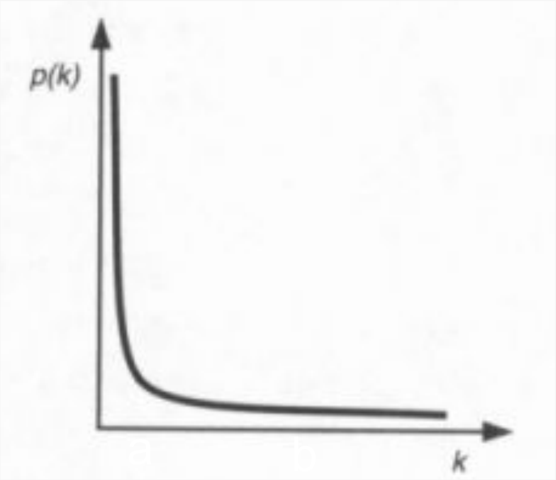
→also called "Heavy tailed" distribution

Typically 2<r<3. For web graph
r ~ 2.1 for in degree distribution
2.7 for out degree distribution

Note that poisson decays *exponentially while power law decays polynomially*

PROPERTIES OF POWER LAW DISTRIBUTIONS

- Ratio of area under the curve [from b to infinity] to [from a to infinity] $=(b/a)^{1-r}$
  - Depends only on the ratio of b to a and not on the absolute values
  - "scale-free"/ "self-similar"
- A moment of order m exists only if $r>m+1$



POWER LAWS

Albert and Barabasi (1999)

Power-law distributions are straight lines in log-log scale.

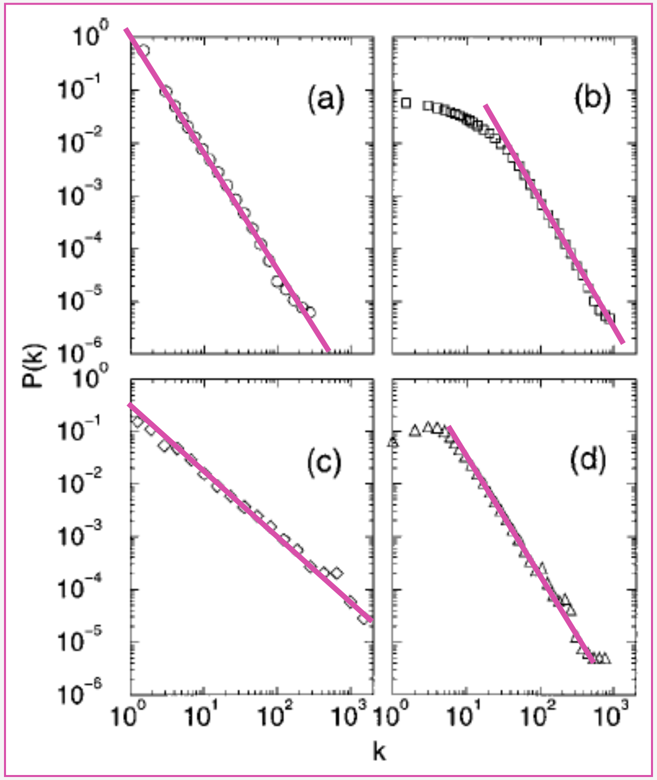
-- slope being r

$y=k^{-r}$ → log y = -r log k → ly= -r lk

How should random graphs be generated to create a power-law distribution of node degrees?

Hint:

Pareto's* Law: Wealth distribution follows a power law.



Power laws in real networks:
(a) WWW hyperlinks
(b) co-starring in movies
(c) co-authorship of physicists
(d) co-authorship of neuroscientists

* Same Velfredo Pareto, who defined Pareto optimality in game theory.

ZIPF'S LAW: POWER LAW DISTRIBUTION BETWEEN RANK AND FREQUENCY

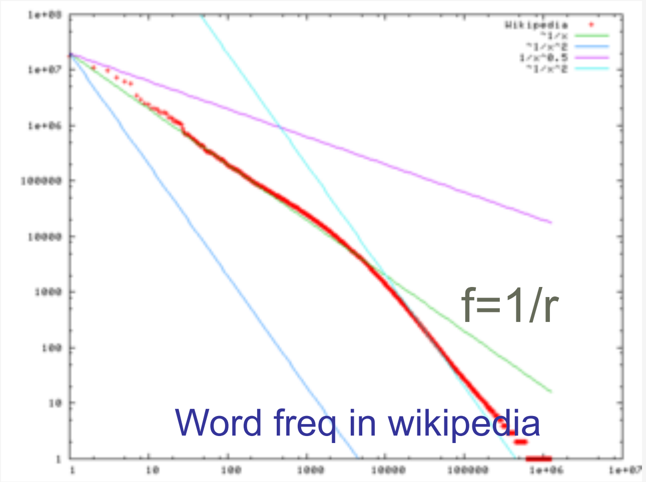
- In a given language corpus, what is the approximate relation between the frequency of a k^{th} most frequent word and $(k+1)^{th}$ most frequent word?

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}$$

For s>1 $\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} < \infty.$



Most popular word is twice as frequent as the second most popular word!

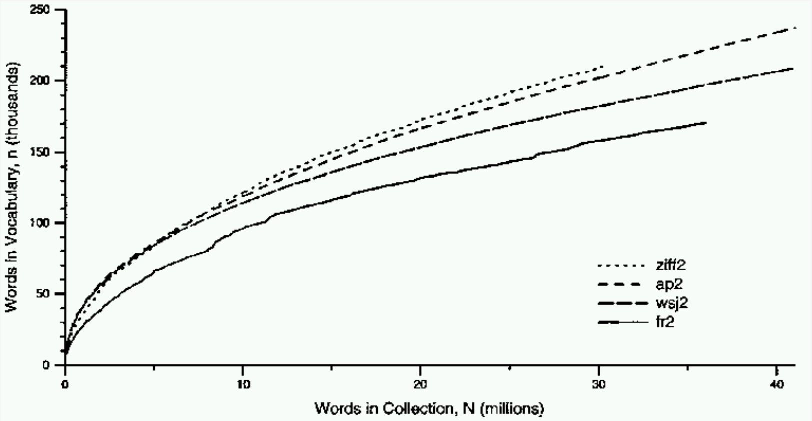
Law of categories in Marketing…

WHAT IS THE EXPLANATION FOR ZIPF'S LAW?

- Zipf's law is an empirical law in that it is observed rather than proved.
- Many explanations have been advanced as to why this holds.
- Zipf's own explanation was "principle of least effort"
  - Balance between speaker's desire for a small vocabulary and hearer's desire for a large one (so meaning can be easily disambiguated)
- Alternate explanation— "rich get richer" –popular words get used more often
- Li (1992) shows that just random typing of letters with space will lead to a "language" with Zipfian distribution..

HEAP'S LAW: A COROLLARY OF ZIPF'S LAW

- What is the relation between the size of a corpus (in terms of words) and the size of the lexicon (vocabulary)?
  - $V = K n^b$
  - K ~ 10—100
  - b ~ 0.4 – 0.6
    - So vocabulary grows as a square root of the corpus size..



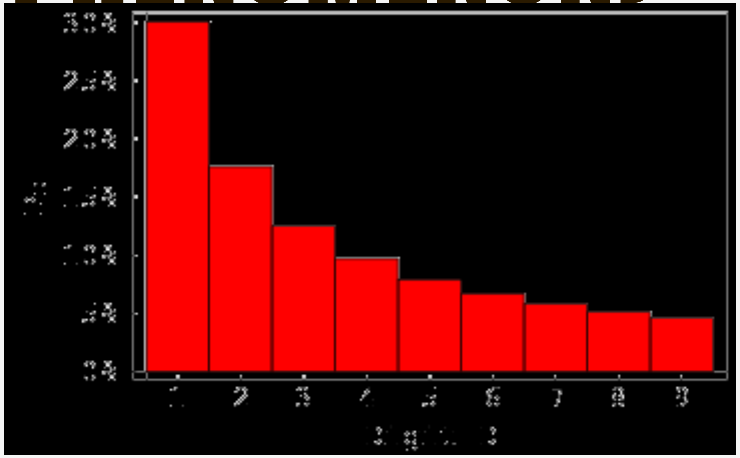
Explanation?
--Assume that the corpus is generated by randomly picking words from a zipfian distribution..

Notice the impact of Zipf on generating random text corpuses!

BENFORD'S LAW (AKA FIRST DIGIT PHENOMENON)

How often does the digit 1 appear in numerical data describing natural phenomenon?

- You would expect 1/9 or 11%



This law holds so well in practice that it is used to catch forged data!!

WHY?

Iff there exists a universal distribution, it must be scale invariant (i.e., should work in any units)

→ starting from there we can show that the distribution must satisfy the differential eqn

$x P'(x) = -P(x)$

For which, the solution is $P(x)=1/x$!

D	P_D	D	P_D
1	0.30103	6	0.0669468
2	0.176091	7	0.0579919
3	0.124939	8	0.0511525
4	0.09691	9	0.0457575
5	0.0791812		

http://mathworld.wolfram.com/BenfordsLaw.html

Digression begets its own digression

POWER LAWS & SCALE-FREE NETWORKS

"The rich get richer!"

Examples of Scale-free networks (i.e., those that exhibit power law distribution of in degree)

- Social networks, including collaboration networks. An example that have been studied extensively is the collaboration of movie actors in films.
- Protein-interaction networks.
- Sexual partners in humans, which affects the dispersal of sexually transmitted diseases.
- Many kinds of computer networks, including the World Wide Web.

Power-law distribution of node-degree arises if

(but *not* "only if")

- As Number of nodes grow edges are added in proportion to the number of edges a node already has.
  - Alternative: Copy model—where the new node copies a random subset of the links of an existing node
    - Sort of close to the WEB reality

SCALE-FREE NETWORKS

- Scale-free networks also exhibit small-world phenomena
  - For a random graph having the same power law distribution as the Web graph, it has been shown that
    - Average path length = $0.35 + \log_{10} N$
- However, scale-free networks tend to be more brittle
  - You can drastically reduce the connectivity by deliberately taking out a few nodes
- This can also be seen as an opportunity..
  - Disease prevention by quarantaining super-spreaders
    - As they actually did to poor Typhoid Mary..

ATTACKS VS. DISRUPTIONS ON SCALE-FREE VS. RANDOM NETWORKS

- Disruption
  - A random percentage of the nodes are removed
- How does the diameter change?
  - Increases monotonically and linearly in random graphs
  - Remains almost the same in scale-free networks
    - Since a random sample is unlikely to pick the high-degree nodes

- Attack
  - A percentage of nodes are removed willfully (e.g. in decreasing order of connectivity)
- How does the diameter change?
  - For random networks, essentially no difference from disruption
    - All nodes are approximately same
  - For scale-free networks, diameter doubles for every 5% node removal!
    - This is an opportunity when you are fighting to contain spread…

EXPLOITING/NAVIGATING SMALL-WORLDS

How does a node in a social network find a path to another node?
→ 6 degrees of separation will lead to n^6 search space (n=num neighbors)
→Easy if we have global graph.. But hard otherwise

- Case 1: Centralized access to network structure
  - Paths between nodes can be computed by shortest path algorithms
    - E.g. All pairs shortest path
  - ..so, small-world ness is trivial to exploit..
    - This is what ORKUT, Friendster etc are trying to do..

- Case 2: Local access to network structure
  - Each node only knows *its own* neighborhood
  - Search without children-generation function ☹
  - Idea 1: Broadcast method
    - Obviously crazy as it increases traffic everywhere
  - Idea 2: Directed search
    - But which neighbors to select?

Are there conditions under which decentralized search can still be easy?

There are very few "fully decentralized" search applications. You normally have hybrid methods between Case 1 and Case 2

Computing one's Erdos number used to take days in the past!

SEARCHABILITY IN SMALL WORLD NETWORKS

- Searchability is measured in terms of Expected time to go from a random source to a random destination
  - We know that in Smallworld networks, the diameter is exponentially smaller than the size of the network.
  - If the expected time is proportional to some small power of log N, we are doing well

Qn: Is this always the case in small world networks?

- To begin to answer this we need to look generative models that take a notion of absolute (lattice or coordinate-based) neighborhood into account

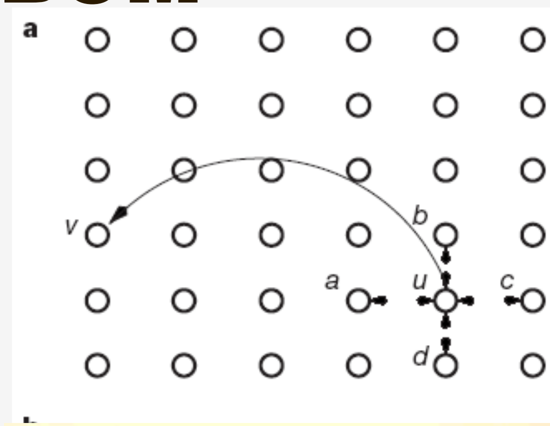
Kleinberg experimented with Lattice networks (where the network is embedded in a lattice—with most connections to the lattice neighbors, but a few shortcuts to distant neighbors)

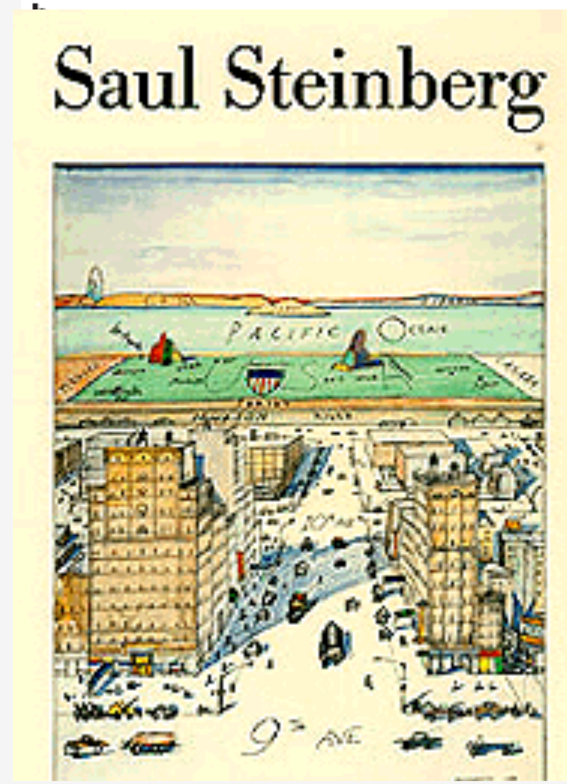
and found that the answer is “Not always”

Kleinberg (2000)

NEIGHBORHOOD BASED RANDOM NETWORKS

- Lattice is *d*-dimensional (*d*=2).
- One random link per node.
- Probability that there is a link between two nodes u and v is $r(u,v)^{-\alpha}$
  - r(u,v) is the "lattice" distance between u and v (computed as manhattan distance)
    - As against geodesic or network distance computed in terms of number of edges
      - E.g. North-Rim and South-Rim
  - α determines how steeply the probability of links to far away neighbors reduces

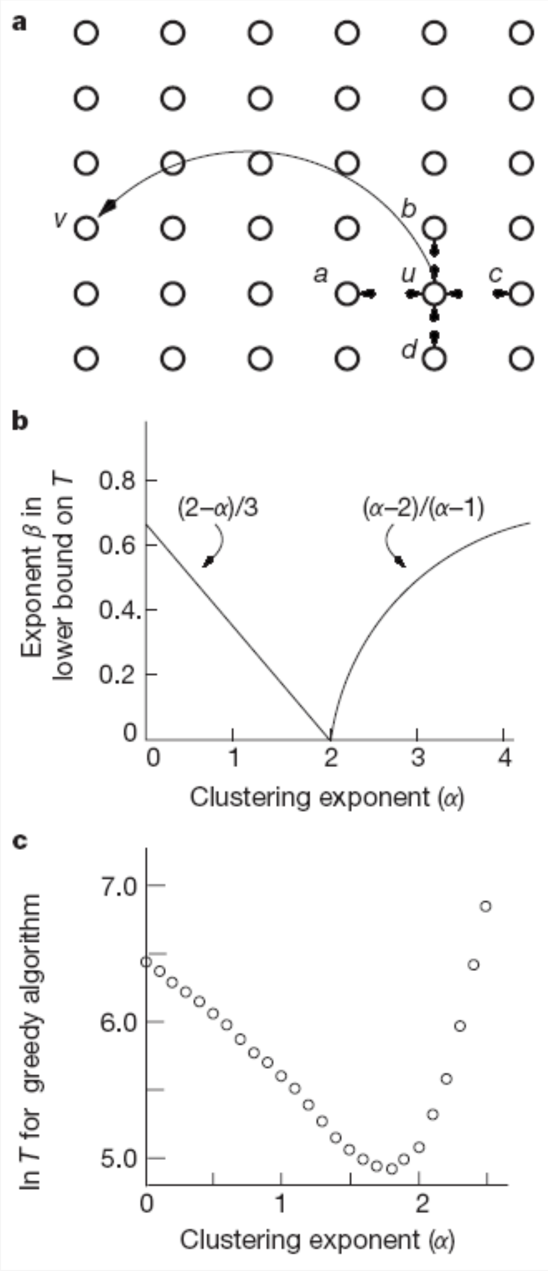




View of the world from 9th Ave

SEARCHEABILITY IN LATTICE NETWORKS

- For d=2, dip in time-to-search at α=2
  – For low α, random graph; no "geographic" correlation in links
  – For high α, not a small world; no short paths to be found.
- Searcheability dips at α=2 (inverse square distribution), in simulation
  - Corresponds to using greedy heuristic of sending message to the node with the least lattice distance to goal
- For d-dimensional lattice, minimum occurs at α=d



SEARCHABLE NETWORKS

Kleinberg (2000)

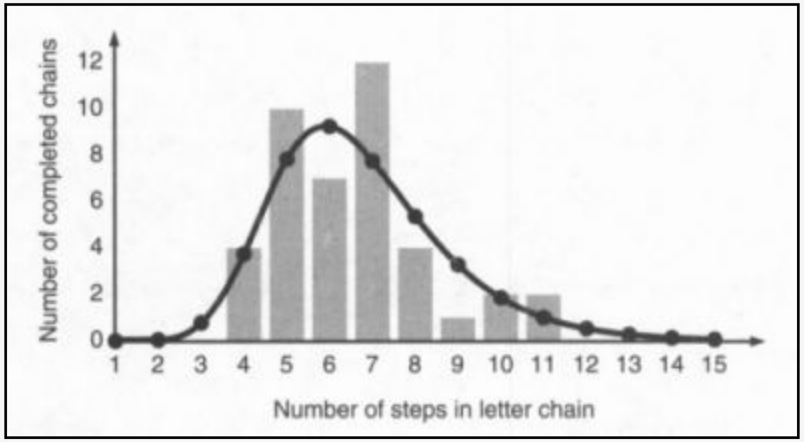
Ramin Zabih

Kentaro Toyama

Watts, Dodds, Newman (2002) show that for *d* = 2 or 3, real networks are quite searchable.

→the dimensions are things like "geography", "profession", "hobbies"

Killworth and Bernard (1978) found that people tended to search their networks by *d* = 2: geography and profession.



The Watts-Dodds-Newman model closely fitting a real-world experiment

DIDN'T MILGRAM'S LETTER EXPERIMENT SHOW THAT NAVIGATION IS EASY?

- …may be not
  - A large fraction of his test subjects *were* stockbrokers
    - So are likely to know how to reach the "goal" stockbroker
  - A large fraction of his test subjects *were* in boston
    - As was the "goal" stockbroker
  - A large fraction of letters never reached
    - Only 20% reached
- So how about (re)doing Milgram experiment with emails?
  - People are even more burned out with (e)mails now
    - Success rate for chain completion < 1% !

NEIGHBORHOOD BASED GENERATIVE MODELS

THESE ESSENTIALLY GIVE MORE LINKS TO CLOSE NEIGHBORS..

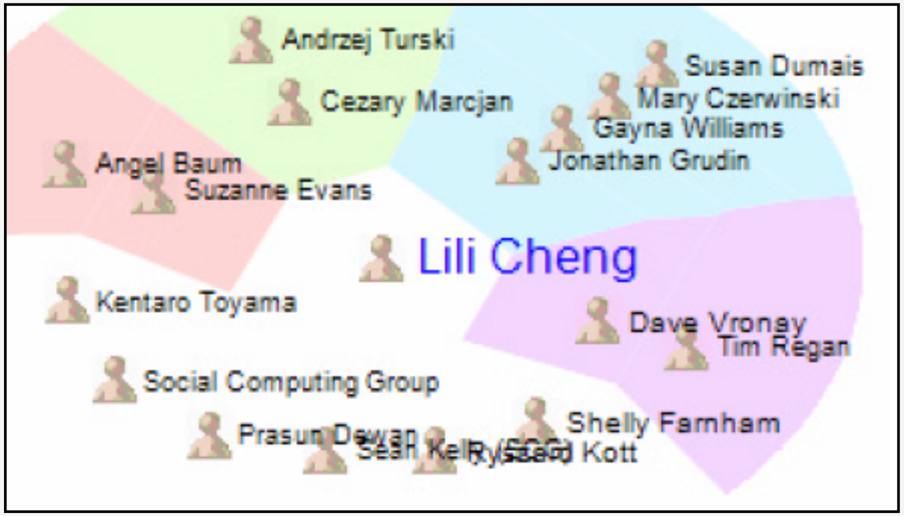
THE ALPHA MODEL

Watts (1999)

The people you know aren't randomly chosen.

People tend to get to know those who are two links away (Rapoport *, 1957).

The real world exhibits a lot of *clustering*.



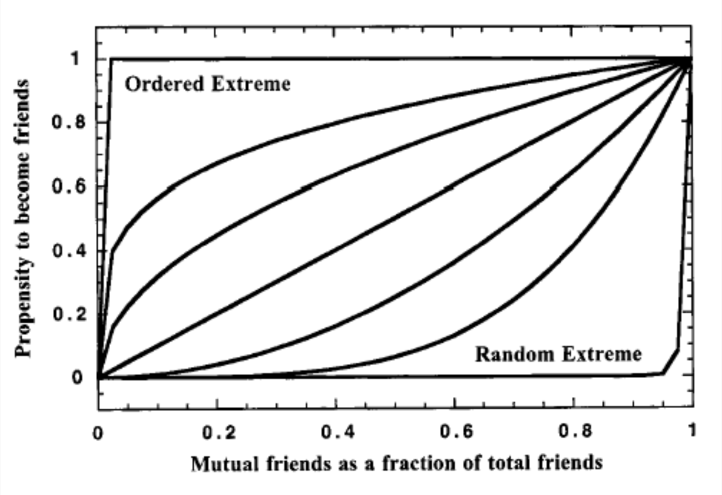
The Personal Map
by MSR Redmond's Social Computing Group

* Same Anatol Rapoport, known for TIT FOR TAT!

THE ALPHA MODEL

Watts (1999)

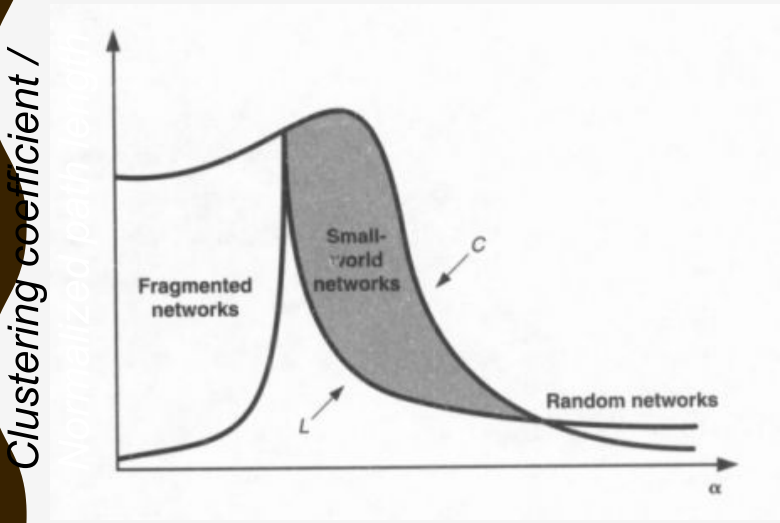
α model: Add edges to nodes, as in random graphs, but makes links more likely when two nodes have a common friend.



Probability of linkage as a function of number of mutual friends
(α is 0 in upper left,
1 in diagonal,
and ∞ in bottom right curves.)

THE ALPHA MODEL

Watts (1999)



Clustering coefficient (C) and average path length (L) plotted against α

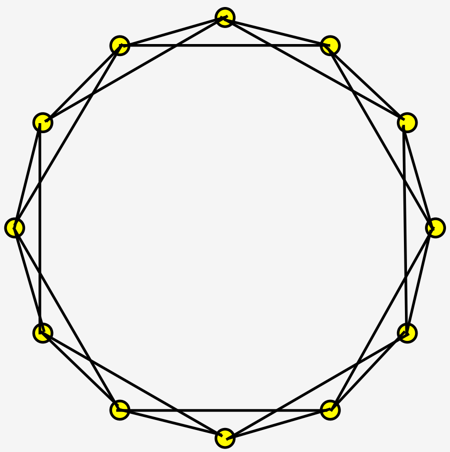
α model: Add edges to nodes, as in random graphs, but makes links more likely when two nodes have a common friend.

For a range of α values:

- The world is small (average path length is short), and
- Groups tend to form (high clustering coefficient).

THE BETA MODEL

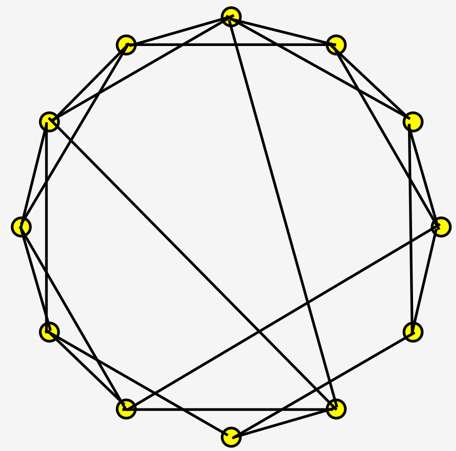
Watts and Strogatz (1998)



$\beta = 0$

People know
their neighbors.

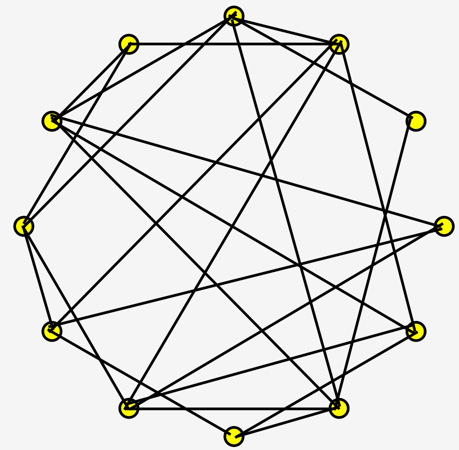
Clustered, but
not a "small world"



$\beta = 0.125$

People know
their neighbors,
and a few distant people.

Clustered and
"small world"



$\beta = 1$

People know
others at
random.

Not clustered,
but "small world"

THE BETA MODEL

Watts and Strogatz (1998)

Nobuyuki Hanaki

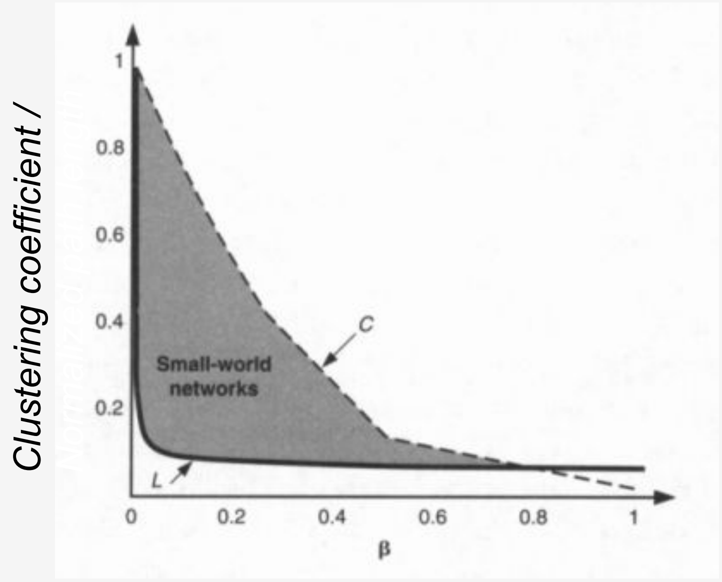
Jonathan Donner

Kentaro Toyama

- First five random links reduce the average path length of the network by half, regardless of *N*!

- Both α and β models reproduce short-path results of random graphs, but also allow for clustering.

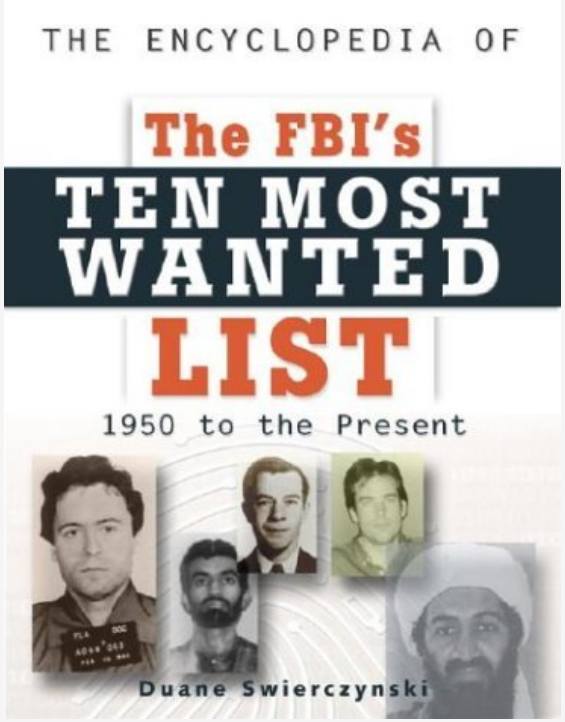
- Small-world phenomena occur at threshold between order and chaos.



Clustering coefficient (C) and average path length (L) plotted against β

SEARCHABLE NETWORKS

Kleinberg (2000)



Just because a short path exists, doesn't mean you can easily find it.

You don't know all of the people whom your friends know.

Under what conditions is a network *searchable*?

SUMMARY

- A network is considered to exhibit small world phenomenon, if its diameter is approximately the logarithm of its size (in terms of number of nodes)
- Most uniformly random networks exhibit small world phenomena
- Most real-world networks are <u>not</u> uniformly random
  - Their in-degree distribution exhibits power-law behavior
  - However, most power-law random networks also exhibit small world phenomena
  - But they are brittle against attack
- The fact that a network exhibits small world phenomenon doesn't mean that an agent with strictly local knowledge can efficiently navigate it (i.e, find paths of O(log(n)) length
  - It is always possible to find the short paths if we have global knowledge
    - This is the case in the FOAF (friend of a friend) networks on the web

REFERENCES

- Textbooks
  - M.E.J. Newman, *Networks: An Introduction*.
  - Albert-Laszlo Barabasi, *Network Science*.
  - David Easley and Jon Kleinberg, *Networks, Crowds and Markets*.
- Popular science books
  - Albert-Laszlo Barabasi, *Linked*.
  - Duncan Watts, *Six Degrees*.
  - Mark Buchanan, *Nexus*.
  - Nicholas Christakis and James Fowler, *Connected*.