Computational Social Science

Course #04199, module 04IN2042

Markus Strohmaier

Prof. Dr.

WeST / IWVI U. of Koblenz-Landau, Germany

e-mail: markus.strohmaier@gesis.org

web: http://www.markusstrohmaier.info

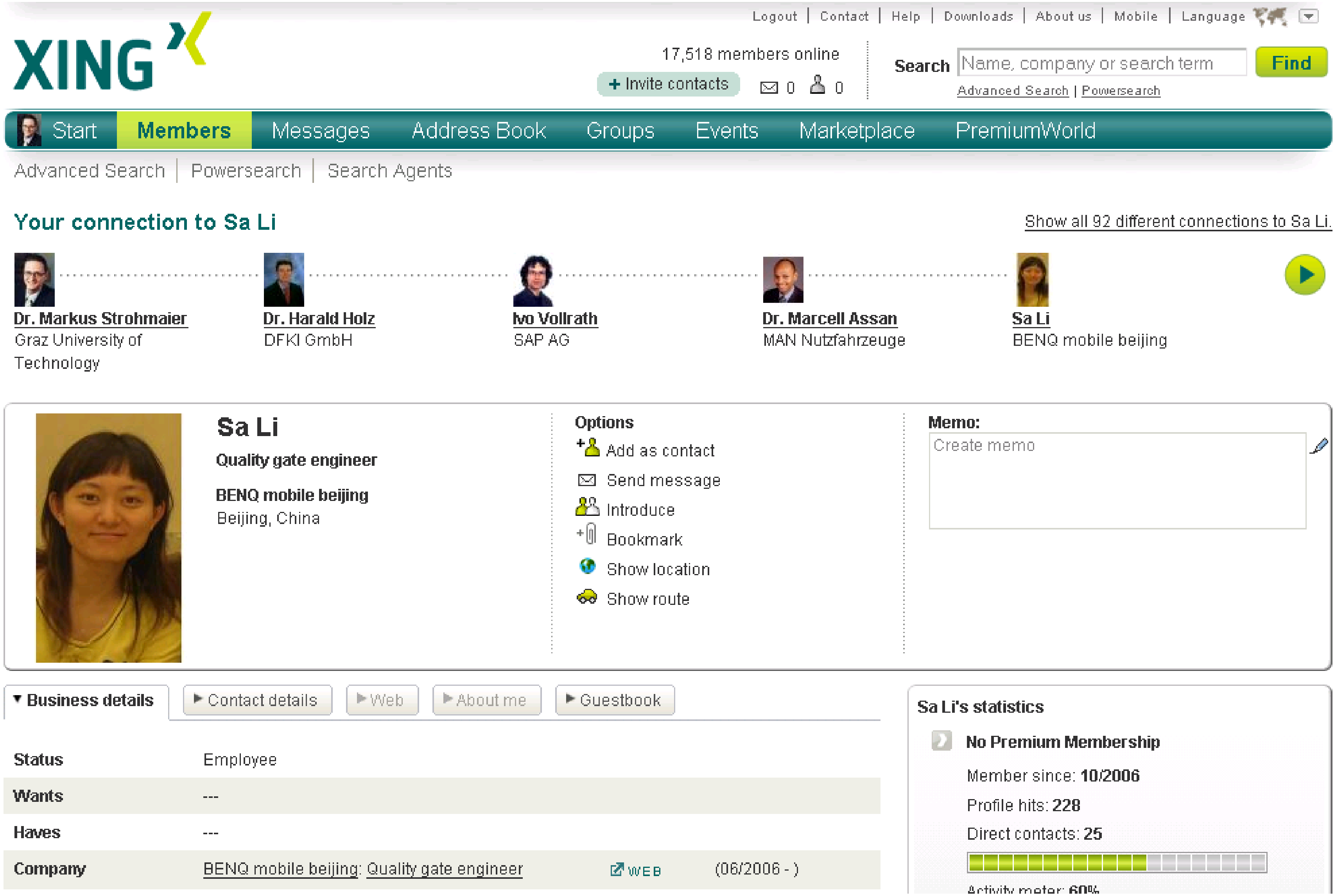
Overview

Topics

- Definition of the Small World Problem
- Results from a social experiment

"every person on earth is connected to any other person through a chain of acquaintances not longer than 6"?

Do I know somebody in …?



M.

The Bacon Number

http://www.imdb.com/

The Kevin Bacon Game

The oracle of Bacon

www.oracleofbacon.org

The Bacon Number
[Watts 2002]

TABLE 3.1 DISTRUBUTION OF ACTORS ACCORDING TO BACON NUMBER

BACON NUMBER	NUMBER OF ACTORS	CUMULATIVE TOTAL NUMBER OF ACTORS
0	1	1
1	1,550	1,551
2	121,661	123,212
3	310,365	433,577
4	71,516	504,733
5	5,314	510,047
6	652	510,699
7	90	510,789
8	38	510,827
9	1	510,828
10	1	510,829

The Erdös Number

Who was Erdös?

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

A famous Hungarian Mathematician, 1913-1996

Erdös posed and solved problems in number theory and other areas and founded the field of discrete mathematics.

- 511 co-authors (Erdös number 1)
- ~ 1500 Publications

The Erdös Number

The Erdös Number:

Through how many research collaboration links is an arbitrary scientist connected to Paul Erdös?

What is a research collaboration link?

Per definition: Co-authorship on a scientific paper -> Convenient: Amenable to computational analysis

What is my Erdös Number?

➔ **5**

me -> S. Easterbrook -> A. Finkelstein -> D. Gabbay -> S. Shelah -> P. Erdös

Stanley Milgram

- A social psychologist
- Yale and Harvard University
- Study on the Small World Problem, beyond well defined communities and relations (such as actors, scientists, …)
- Controversial: The Obedience Study
- What we will discuss today:
  „An Experimental Study of the Small World Problem"

1933-1984

Introduction

The simplest way of formulating the small-world problem is:
Starting with any two people in the world, what is the likelihood that they will know each other?

A somewhat more sophisticated formulation, however, takes account of the fact that while person X and Z may not know each other directly, they may share a mutual acquaintance - that is, a person who knows both of them. One can then think of an acquaintance chain with X knowing Y and Y knowing Z.

Moreover, one can imagine circumstances in which X is linked to Z not by a single link, but by a series of links, X-A-B-C-D…Y-Z. That is to say, person X knows person A who in turn knows person B, who knows C… who knows Y, who knows Z.

[Milgram 1967, according to
]http://www.ils.unc.edu/dpr/port/socialnetworking/theory_paper.html#2]

An Experimental Study of the Small World Problem [Travers and Milgram 1969]

A Social Network Experiment tailored towards

- Demonstrating
- Defining
- And measuring

Inter-connectedness in a large society (USA)

A test of the modern idea of “six degrees of separation”

Which states that: **every person on earth is connected to any other person through a chain of acquaintances not longer than 6**

Experiment

Goal

- Define a single target person and a group of starting persons
- Generate an acquaintance chain from each starter to the target

Experimental Set Up

- Each starter receives a document
- was asked to begin moving it by mail toward the target
- Information about the target: name, address, occupation, company, college, year of graduation, wife's name and hometown
- Information about relationship (*friend/acquaintance*) [Granovetter 1973]

Constraints

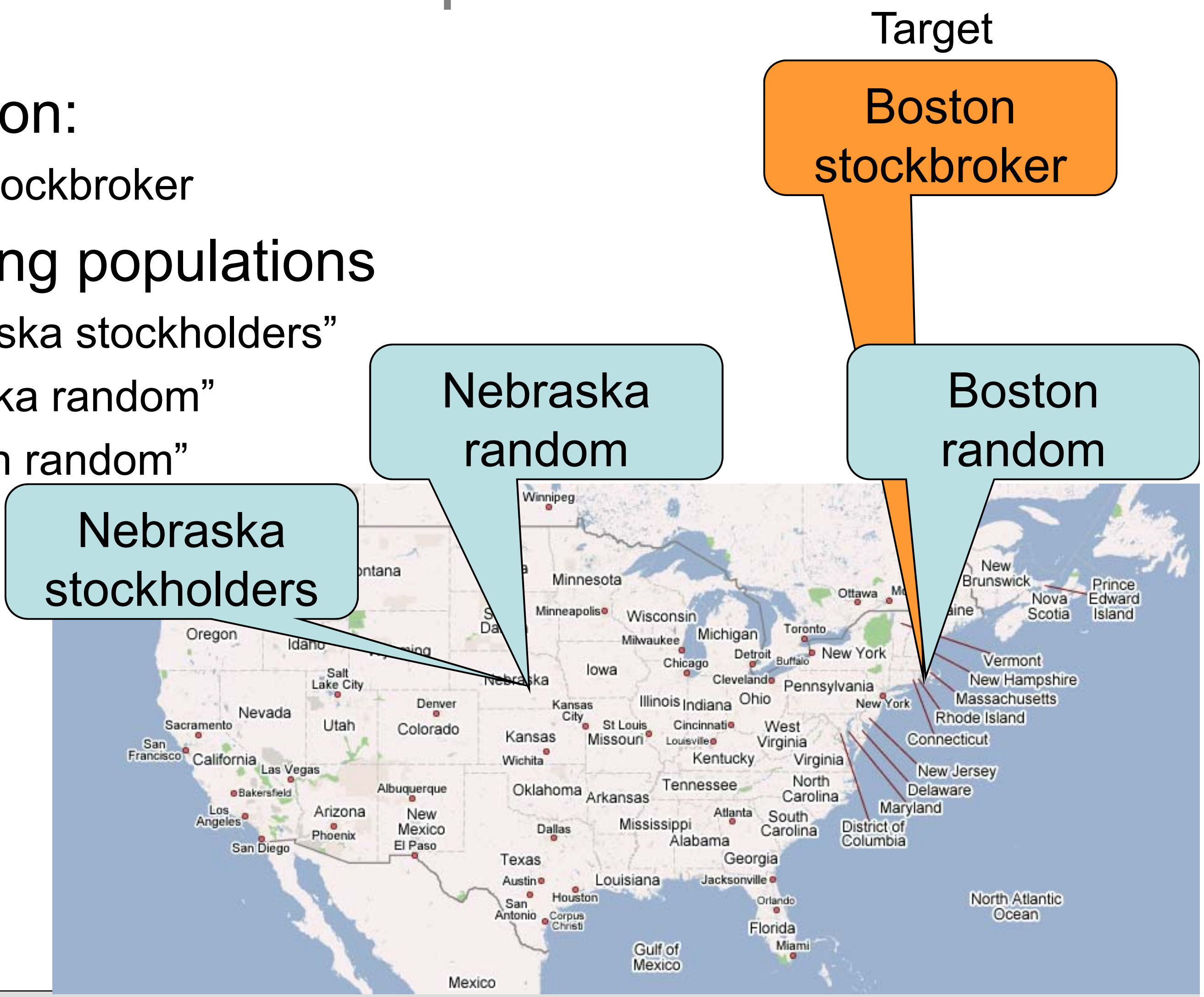
- starter group was only allowed to send the document to people they know and
- was urged to choose the next recipient in a way as to advance the progress of the document toward the target

Questions

- How many of the starters would be able to establish contact with the target?
- How many intermediaries would be required to link starters with the target?
- What form would the distribution of chain lengths take?

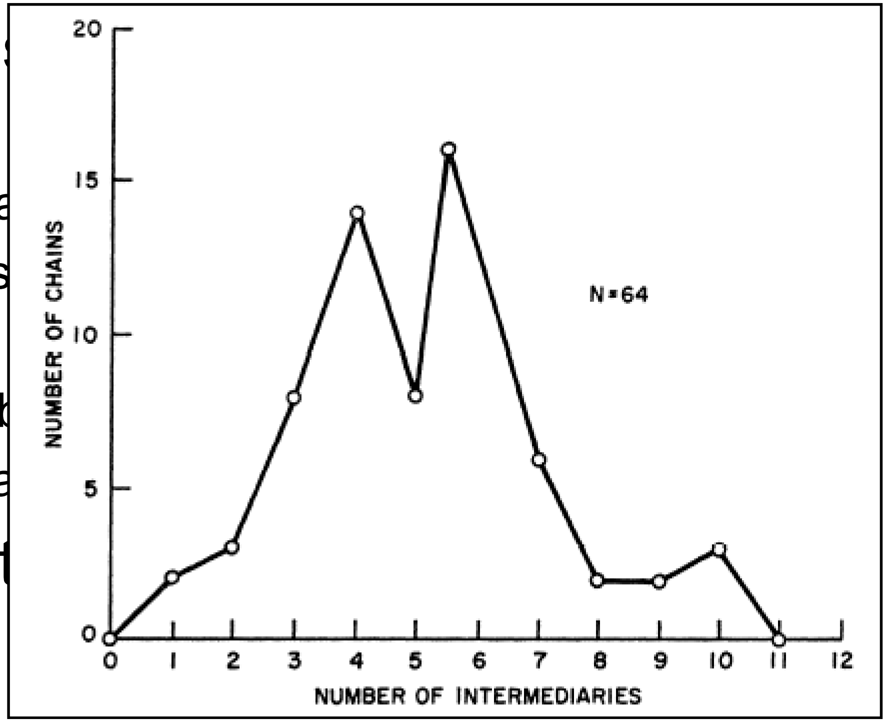
Set Up

- Target person:
  - A Boston stockbroker
- Three starting populations
  - 100 “Nebraska stockholders”
  - 96 “Nebraska random”
  - 100 “Boston random”



Results I

- How many of the starters would be able to establish contact with the target?
  - 64 out of 296 reached the target
- How many intermediaries
  starters with the target?
  - Well, that depends: the overa
  - Through hometown: 6.1 links
  - Through business: 4.6 links
  - Boston group faster than Neb
  - Nebraska stockholders not fa
- What form would the dist
  take?



Results II

- Incomplete chains

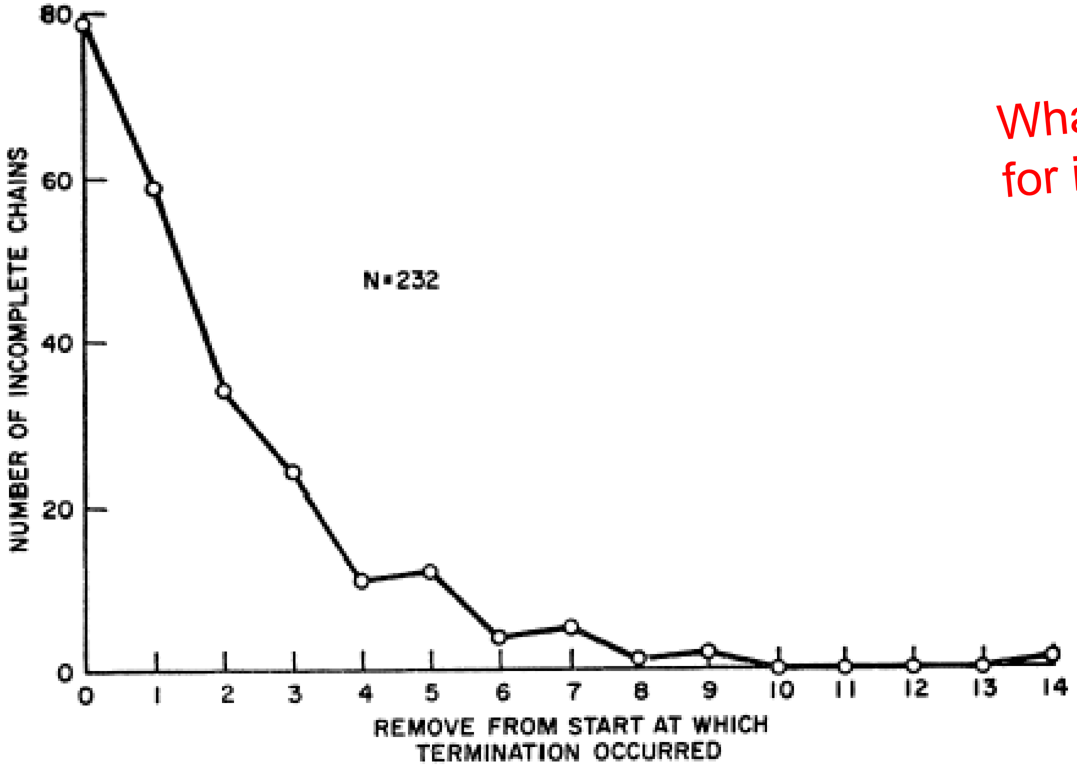


FIGURE 2

Lengths of Incomplete Chains

What reasons can you think of for incomplete chains?

Results III

- Common paths
- Also see:
  Gladwell's "Law of the few"

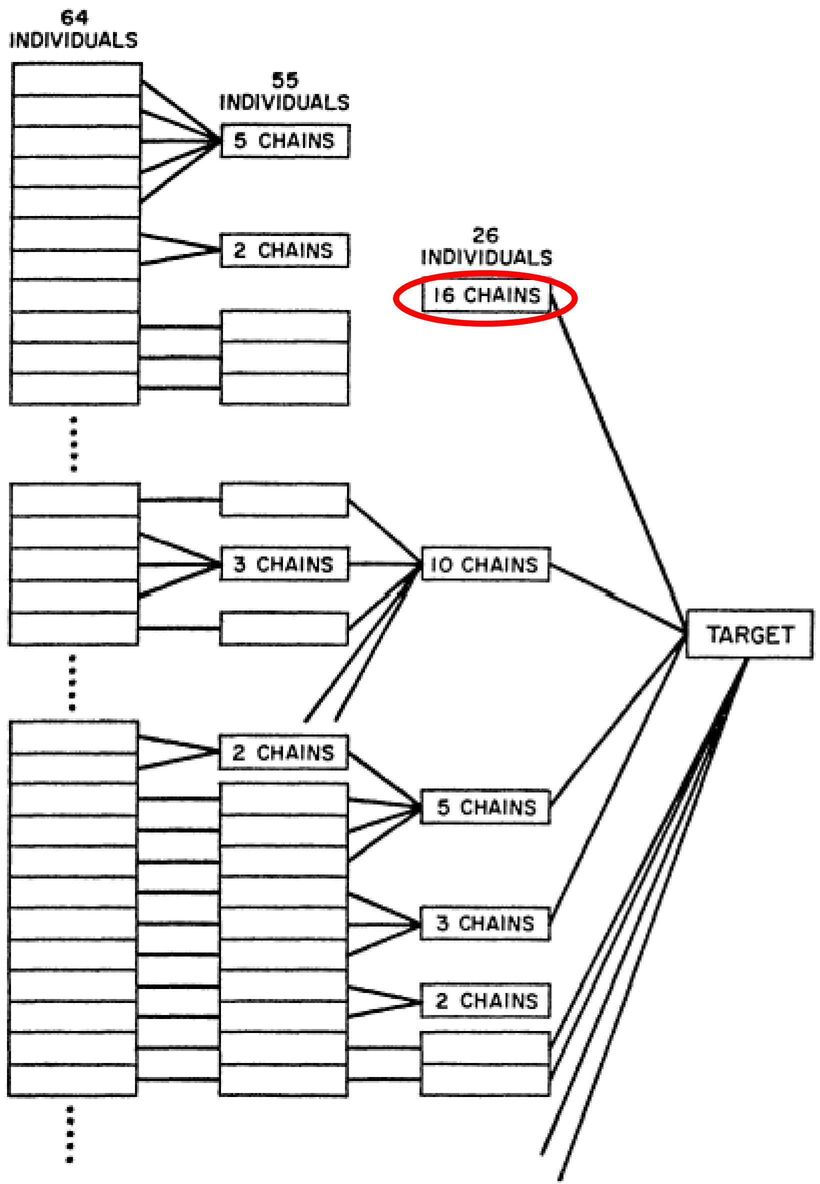


FIGURE 3

Common Paths Appear as Chains Converge on the Target

6 degrees of separation

- So is there an upper bound of six degrees of separation in social networks?

What kind of problems do you see with the results of this study?

  - Extremely hard to test
  - In Milgram's study, ~2/3 of the chains didn't reach the target
  - 1/3 random, 1/3 blue chip owners, 1/3 from Boston
  - Danger of loops (mitigated in Milgram's study through chain records)
  - Target had a "high social status" [Kleinfeld 2000]

Follow up work (2008)

http://arxiv.org/PS_cache/arxiv/pdf/0803/0803.0939v1.pdf

– Horvitz and Leskovec study 2008
– 30 billion conversations among 240 million people of Microsoft Messenger
– Communication graph with 180 million nodes and 1.3 billion undirected edges
– Largest social network constructed and analyzed to date (2008)

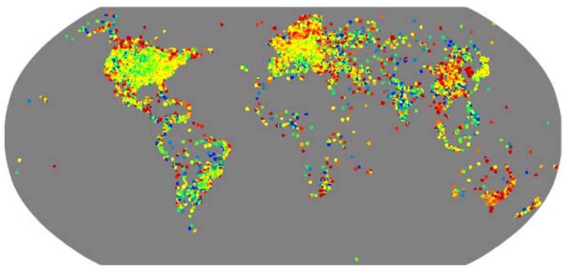


Figure 10: Number of users at a particular geographic location. Color represents the number of users. Notice the map of the world appears.

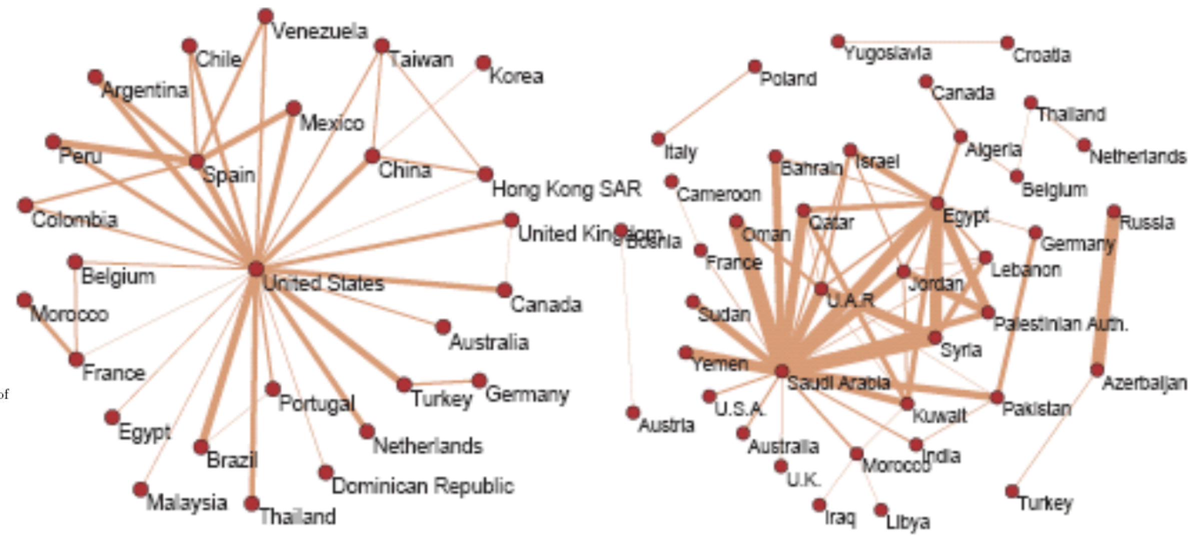
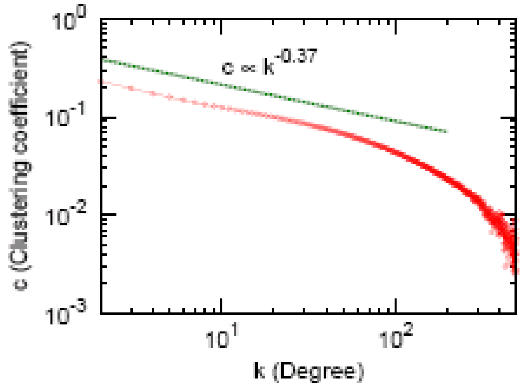


Figure 14: (a) Communication among countries with at least 10 million conversations in June 2006. (b) Countries by average length of the conversation. Edge widths correspond to logarithms of intensity of links.

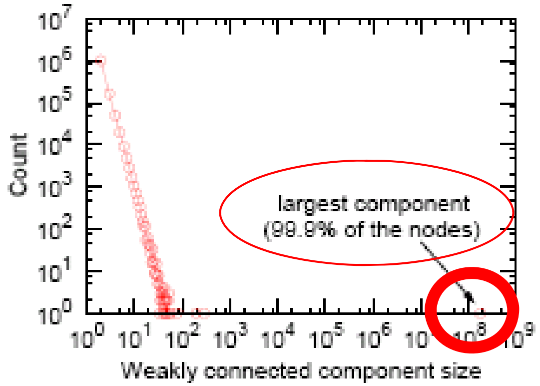
Follow up work (2008)

http://arxiv.org/PS_cache/arxiv/pdf/0803/0803.0939v1.pdf

- the clustering coefficient decays very slowly with exponent −0.37 with the degree of a node and the average clustering coefficient is 0.137.
- This result suggests that clustering in the Messenger network is much higher than expected—that people with common friends also tend to be connected.



(a) Clustering



(b) Components

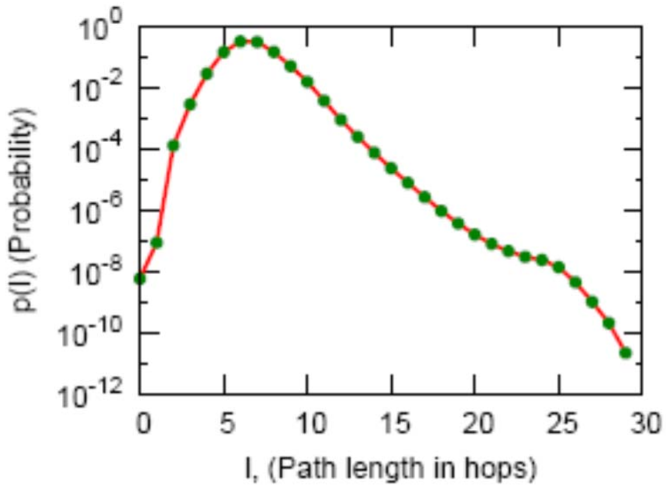
Figure 19: (a) Clustering coefficient; (b) distribution of connected components. 99.9% of the nodes belong to the largest connected component.

Follow up work (2008)

http://arxiv.org/PS_cache/arxiv/pdf/0803/0803.0939v1.pdf

Approximation of “Degrees of separation”

- Random sample of 1000 nodes
- for each node the shortest paths to all other nodes was calculated. The average path length is 6.6. median at 7.
- Result: a random pair of nodes is 6.6 hops apart on the average, which is half a link longer than the length reported by Travers/Milgram.
- The 90th percentile (effective diameter (16)) of the distribution is 7.8. 48% of nodes can be reached within 6 hops and 78% within 7 hops.
- we find that there are about “7 degrees of separation” among people.
- long paths exist in the network; we found paths up to a length of 29.

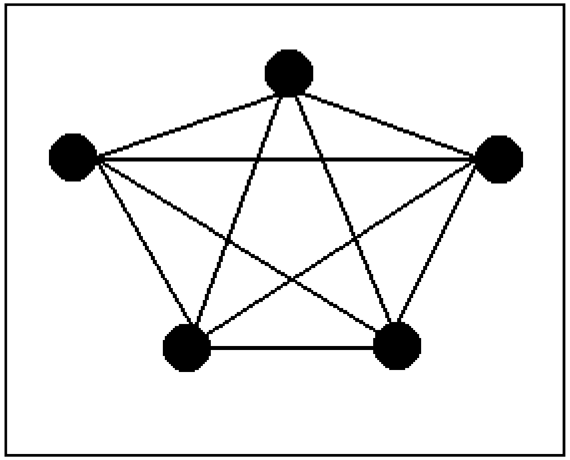


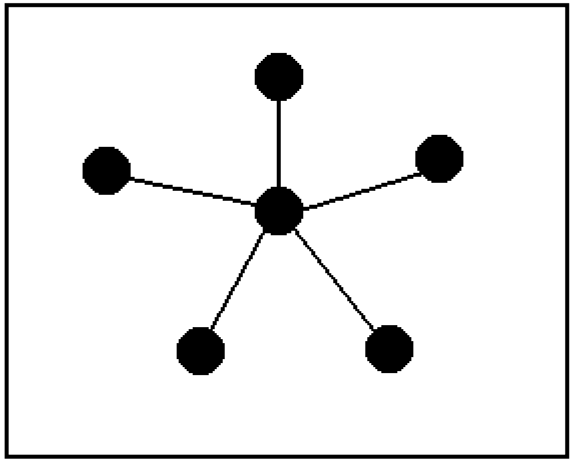
(a) Diameter

Small Worlds

http://www.infosci.cornell.edu/courses/info204/2007sp/

- Every pair of nodes in a graph is connected by a path with an extremely small number of steps
  (low diameter)
- Two principle ways of encountering small worlds
  - Dense networks
  - sparse networks with well-placed connectors

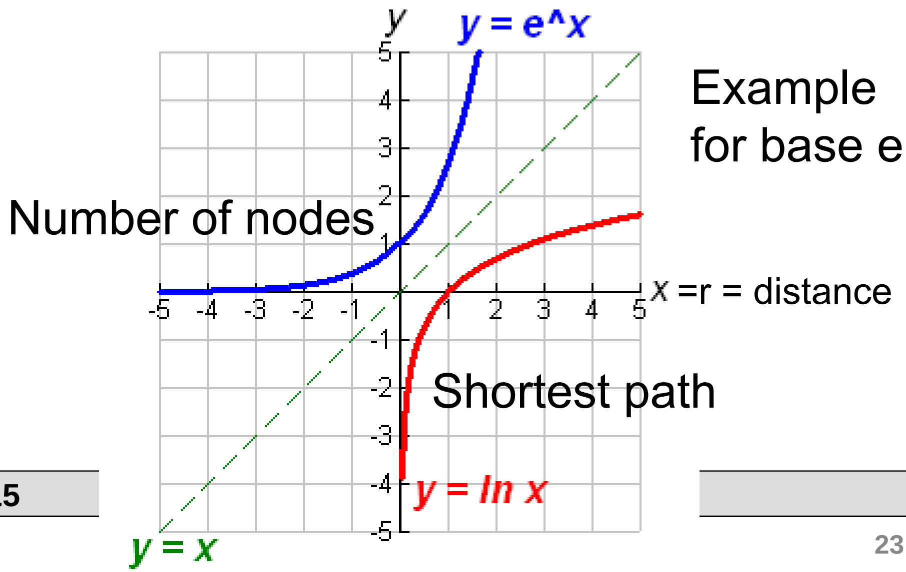




Small Worlds
[Newman 2003]

- The small-world effect exists, if
  - „The number of vertices within a distance r of a typical central vertex grows exponentially with r (the larger it get, the faster it grows) $x(t) = x_0 e^{kt}$

  In other words:
  - Networks are said to show the small-world effect if the value of l (avg. shortest distance) scales logarithmically or slower with network size for fixed mean degree $e^{\ln(x)} = x \quad \text{if } x > 0$



Formalizing the Small World Problem

[Watts and Strogatz 1998]

The small-world phenomenon is assumed to be present when

$L \gtrsim L_{\text{random}}$ but $C >> C_{\text{random}}$

Or in other words: We are looking for networks where local clustering is high and global path lengths are small

What's the rationale for the above formalism?

One potential answer:

Cavemen and Solaris Worlds

The Solaris World
Random Social Connections

How do random social graphs differ
from „real“ social networks?

http://vimeo.com/9669721

http://bits.blogs.nytimes.com/2010/02/13/chatroulettes-founder-17-introduces-himself/

The Cave Men World
Highly Clustered Social Connections

Formalizing the Small World Problem

[Watts 2003]

Reminder - previous informal definition:
SMP exists when every pair of nodes in a graph is connected by a path with an extremely small number of steps.
Does not take searchability into account. Random networks are hard to se

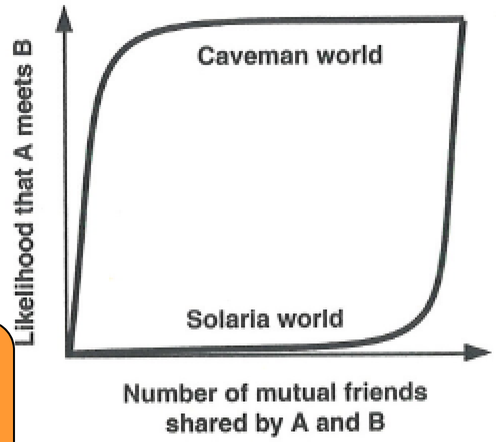


Figure 3.1. Two extreme kinds of interaction rules. In the top curve (caveman world), even a single mutual friend implies that A and B are highly likely to meet. In the bottom curve (Solaria world), all interactions are equally unlikely, regardless of how many friends A and B share.

Under which conditions can these two requirements be reconciled?

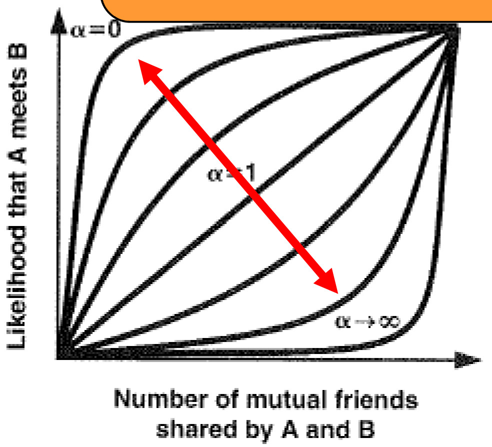


Figure 3.2. Be... two extremes, a whole family of interaction rules exists, each one specified by a particular value of the tuneable parameter alpha (α). When α = 0, we have a caveman world; when α become... we have Solaria.

Two seemingly contradictory requirements for the Small World Phenomenon:

- It should be possible to connect two people chosen at random via chain of only a few intermediaries (as in Solaria world)
- Network should display a large clustering coefficient, so that a node's friends will know each other (as in Caveman world)

Search-ability

Formalizing the Small World Problem

[Watts 2003]

- Page 76 -82
- The alpha parameter
- Path length: computed only over nodes in the same connected component

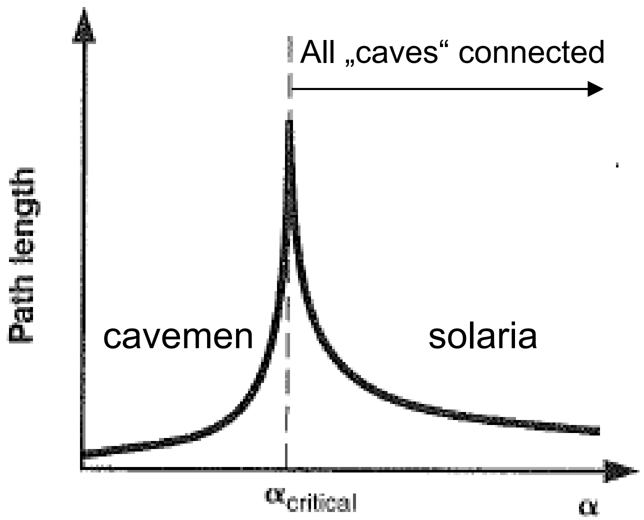
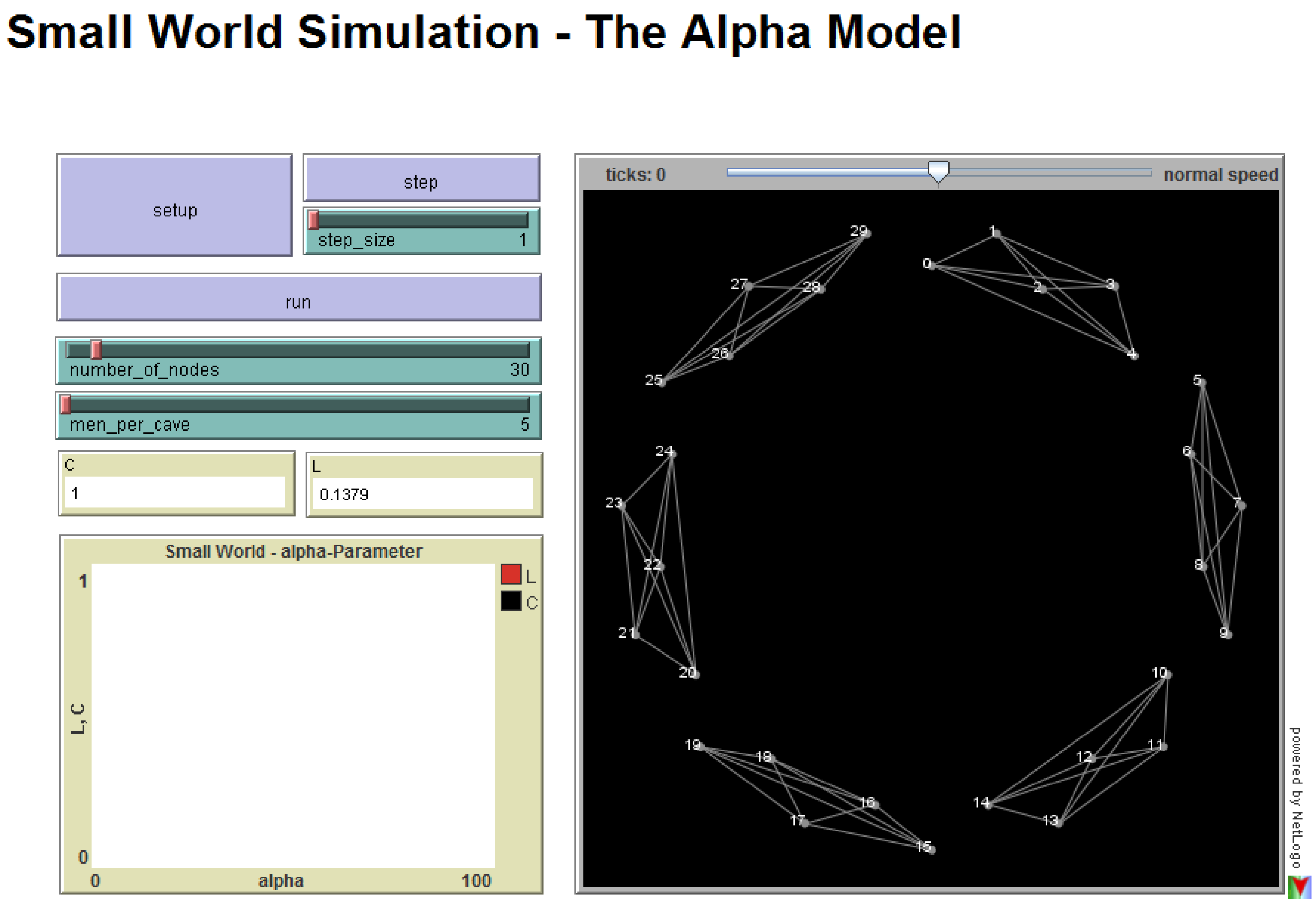


Figure 3.3. Path length as a function of alpha (α). At the critical alpha value, many small clusters join to connect the entire network, whose length then shrinks rapidly.

Demo – Small Worlds the Alpha Model

http://markusstrohmaier.info/demos/sw-alpha.htm



Formalizing the Small World Problem

[Watts 2003]

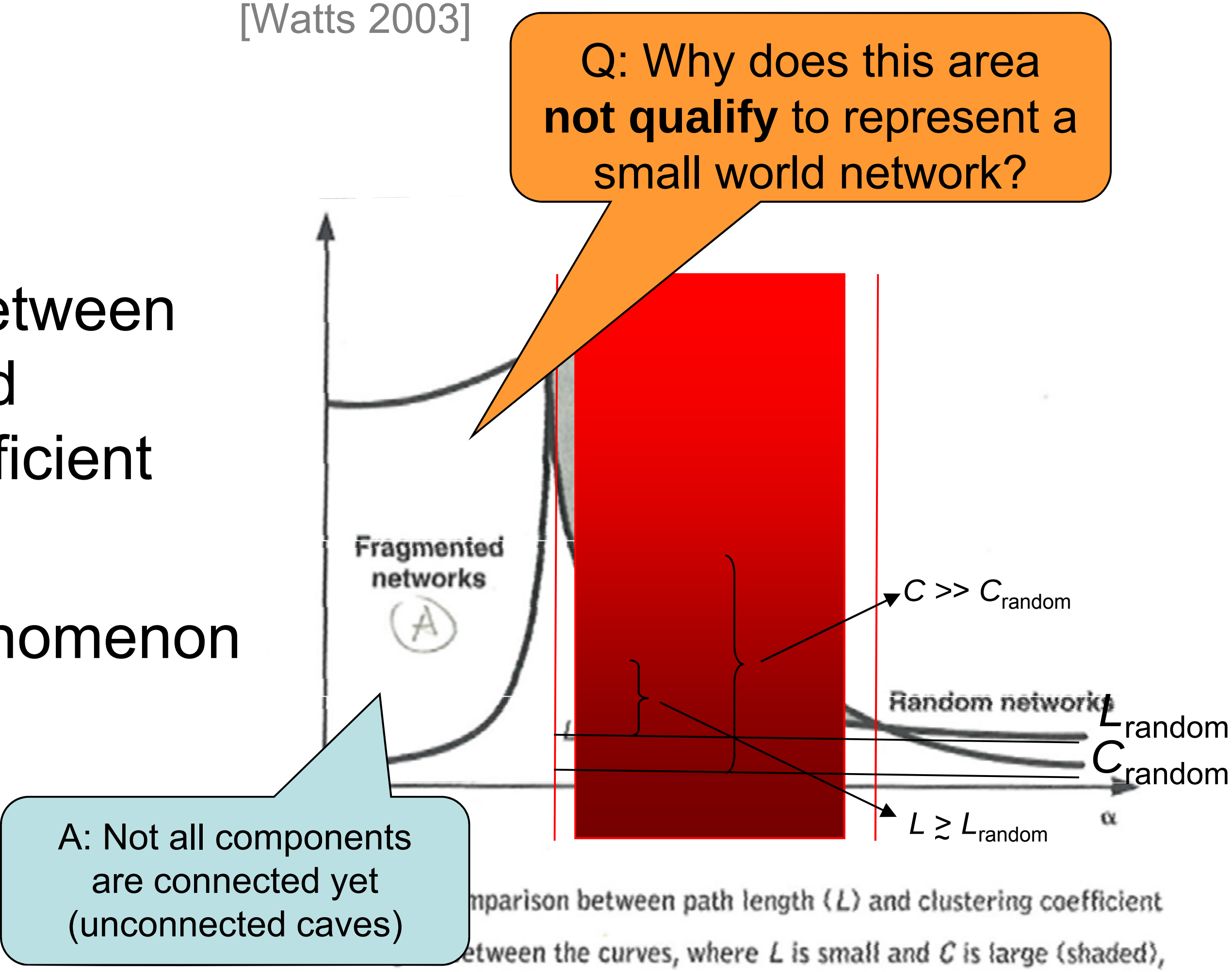
- Page 76 -82

- Comparison between path length and clustering coefficient

Small World Phenomenon exists when

$L \gtrsim L_{random}$ but

$C >> C_{random}$



…mparison between path length (L) and clustering coefficient …etween the curves, where L is small and C is large (shaded), represents the presence of small-world networks.

Examples for Small World Networks

[Watts and Strogatz 1998]

Table 1 Empirical examples of small-world networks

$L > L_{random}$ but $C >> C_{random}$	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

Characteristic path length L and clustering coefficient C for three real networks, compared to random graphs with the same number of vertices (n) and average number of edges per vertex (k). (Actors: $n = 225{,}226$, $k = 61$. Power grid: $n = 4{,}941$, $k = 2.67$. *C. elegans*: $n = 282$, $k = 14$.) The graphs are defined as follows. Two actors are joined by an edge if they have acted in a film together. We restrict attention to the giant connected component[16] of this graph, which includes ~90% of all actors listed in the Internet Movie Database (available at http://us.imdb.com), as of April 1997. For the power grid, vertices represent generators, transformers and substations, and edges represent high-voltage transmission lines between them. For *C. elegans*, an edge joins two neurons if they are connected by either a synapse or a gap junction. We treat all edges as undirected and unweighted, and all vertices as identical, recognizing that these are crude approximations. All three networks show the small-world phenomenon: $L \gtrsim L_{random}$ but $C \gg C_{random}$.

Formalizing the Small World Problem
[Watts 2003]

- Page 87

- The beta

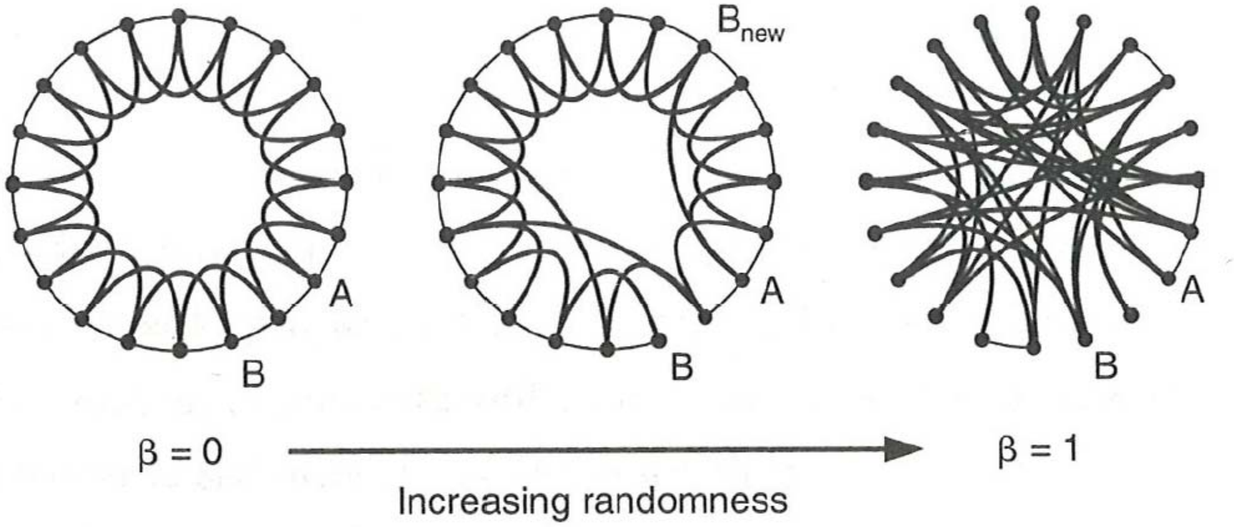


Figure 3.6. Construction of the beta model. The links in a one-dimensional, periodic lattice are randomly rewired with probability beta (β). When beta is zero (left), the lattice remains unchanged, and when beta is one (right), all links are rewired, generating a random network. In the middle, networks are partly ordered and partly random (for example, the original link from A to B has been rewired to B_{new}).

Formalizing the Small World Problem

[Watts 2003]

- Page 87 -90

- The beta parameter

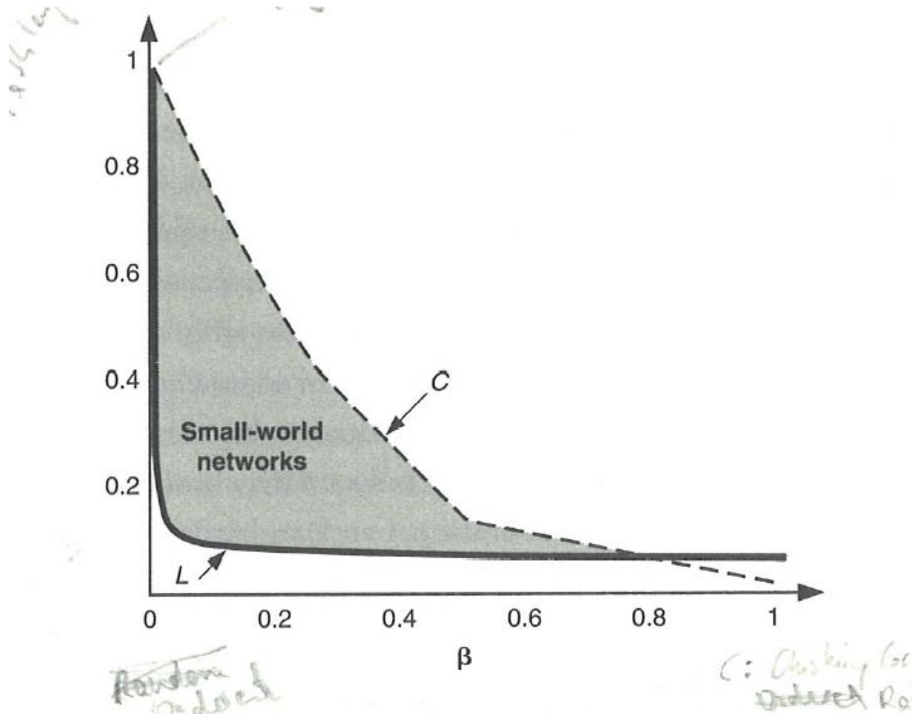
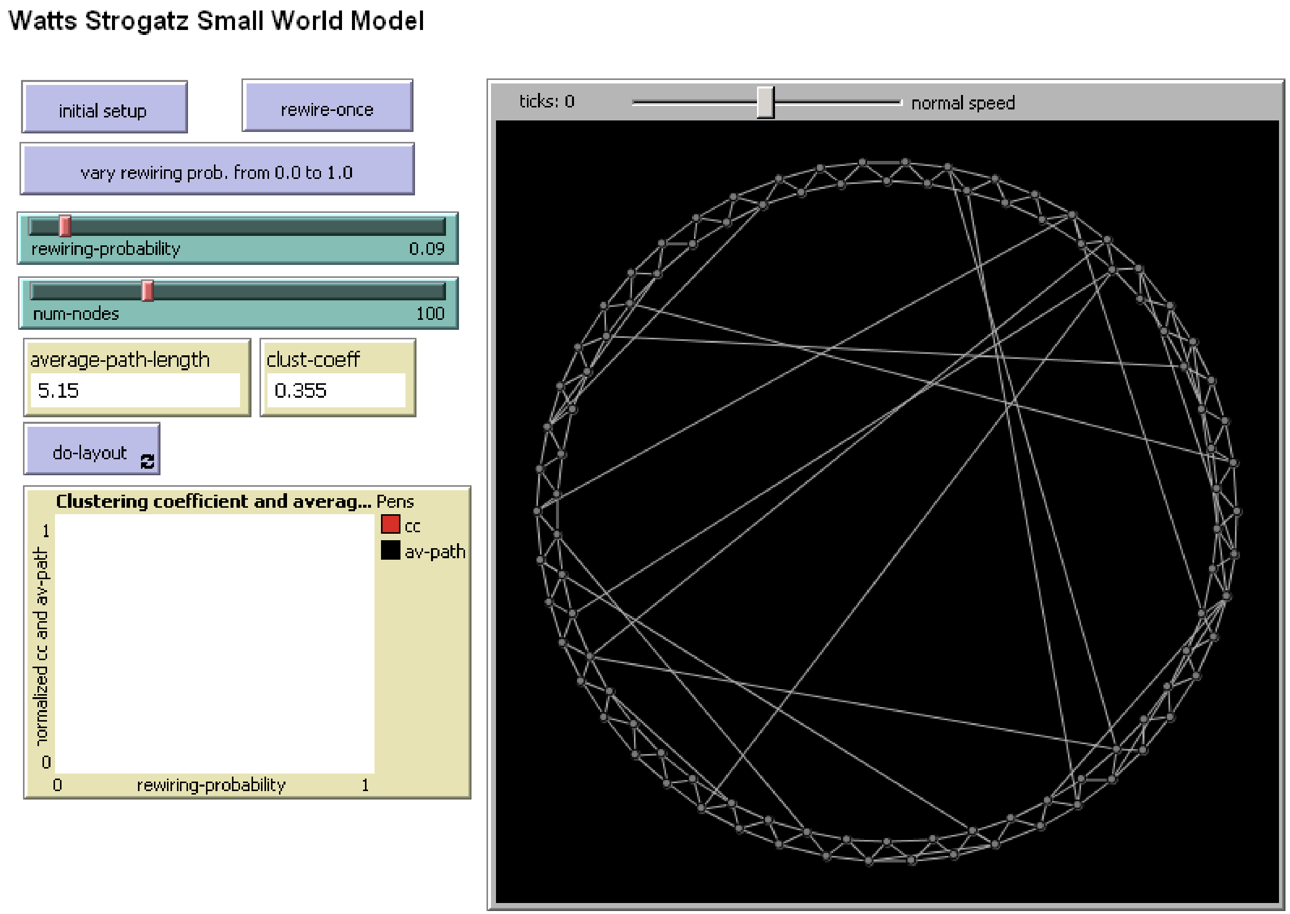


Figure 3.7. Path length and clustering coefficient in the beta model. As with the alpha model (see Figure 3.4), small-world networks exist when path length is small and the clustering coefficient is large (shaded region).

Demo – Small Worlds

http://www.ladamic.com/netlearn/NetLogo4/SmallWorldWS.html



Contemporary Software

- Where does the small-world phenomenon come into play in contemporary software, in organizations, ..?

- Xing, LinkedIn, Myspace, Facebook, FOAF, …
- Business Processes, Information and Knowledge Flow

How do Small World Networks form?

Preferential Attachment [Barabasi 1999]

„The rich getting richer“

Preferential Attachment refers to the high probability of a new vertex to connect to a vertex that already has a large number of connections

Example:

1. a new website linking to more established ones
2. a new individual linking to well-known individuals in a social network

Preferential Attachment Example

Which node has the highest probability of being linked by a new node in a network that exhibits traits of preferential attachment?

Example

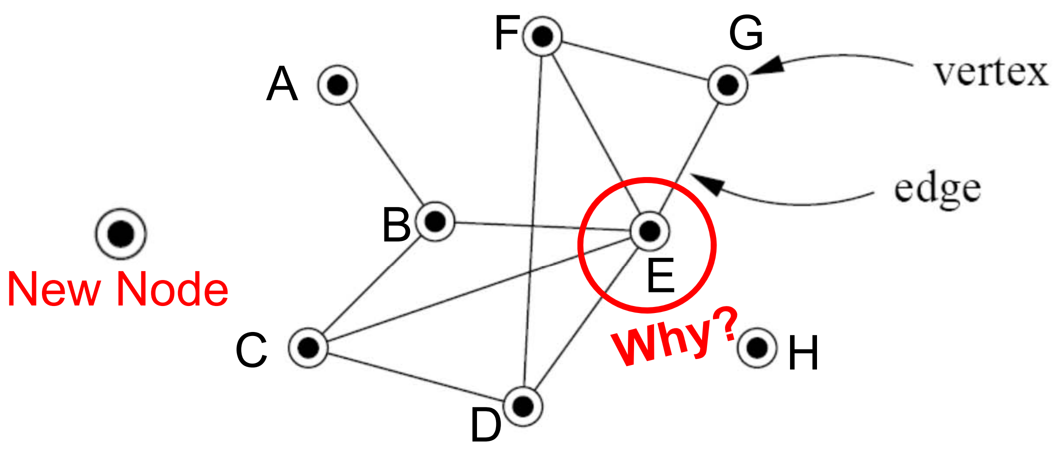


FIG. 1 A small example network with eight vertices and ten edges.

[Newman 2003]

Assortative Mixing (or Homophily) [Newman 2003]

Assortative Mixing refers to selective linking of nodes to other nodes who share some common property

- E.g. degree correlation
  high degree nodes in a network associate preferentially with other high-degree nodes
- E.g. social networks
  nodes of a certain type tend to associate with the same type of nodes (e.g. by race)

Assortative Mixing (or Homophily) [Newman 2003]

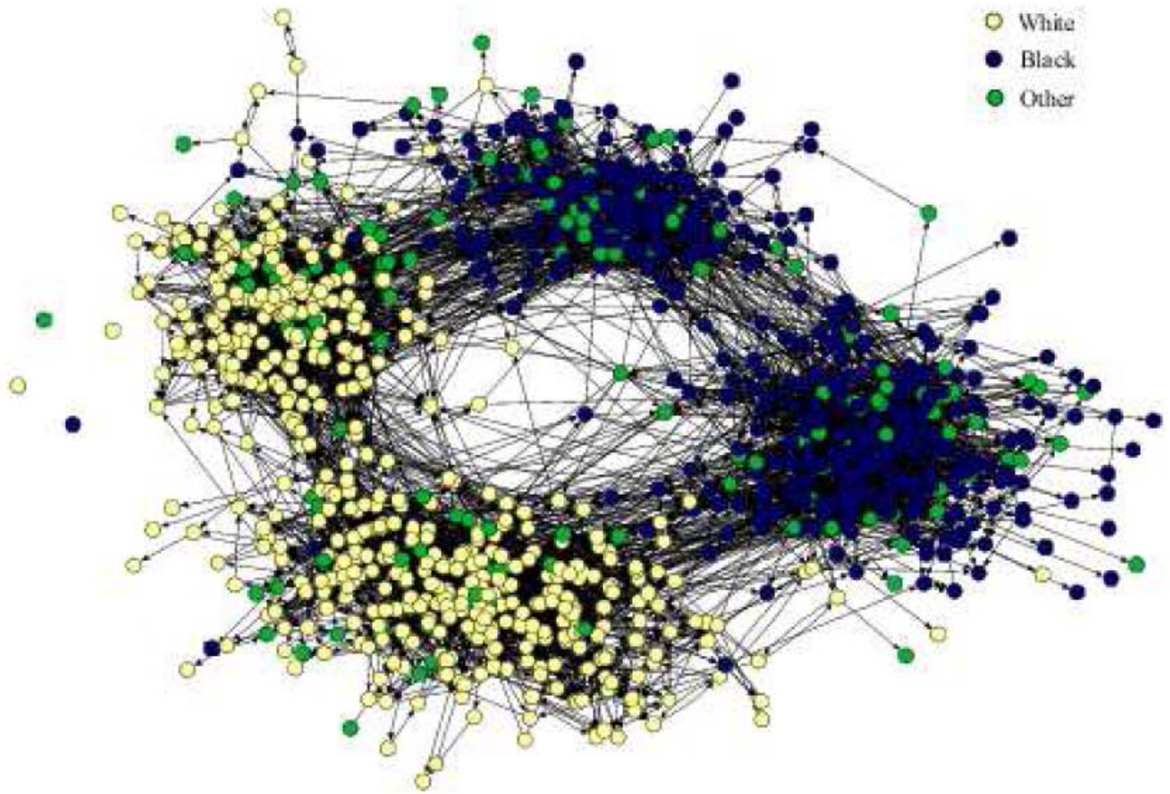


FIG. 8 Friendship network of children in a US school. Friendships are determined by asking the participants, and hence are directed, since A may say that B is their friend but not *vice versa*. Vertices are color coded according to race, as marked, and the split from left to right in the figure is clearly primarily along lines of race. The split from top to bottom is between middle school and high school, i.e., between younger and older children. Picture courtesy of James Moody.

Disassortativity
[Newman 2003]

Disassortativity refers to selective linking of nodes to other nodes who are different in some property

- E.g. the web
  low degree nodes tend to associate with high degree nodes

Any questions?

See you next week!