

Probability & statistics seminar, April 4, 2002.

---

# Identity and Search in Social Networks

---

*D.J. Watts<sup>1,2,3</sup>, M.E.J. Newman<sup>2</sup>, P.S. Dodds<sup>3</sup>*

*1. Department of Sociology, Columbia University*

*2. Santa Fe Institute*

*3. Earth Institute, Columbia University*

Outline:

1. Description of the small-world problem.
2. Review of some previous work.
3. Current model.
4. Applications/conclusions.

The problem:

How are social networks **structured**?

- How do we define connections?

What about the **dynamics** of social networks?

- How do social networks evolve?
- How do social movements begin?
- Why do groups break apart?
- How is information transmitted through social networks?
- Does communication depend more on the network or the senders?

*The problem:*

Can solutions to sociology problems inform other areas of research, even physics?

The problem:

One small slice of the pie:

**Q.** Can people pass messages to each other using only their social connections?

**A.** Yes (apparently):

The small-world phenomenon  
or

“Six Degrees of Separation.”

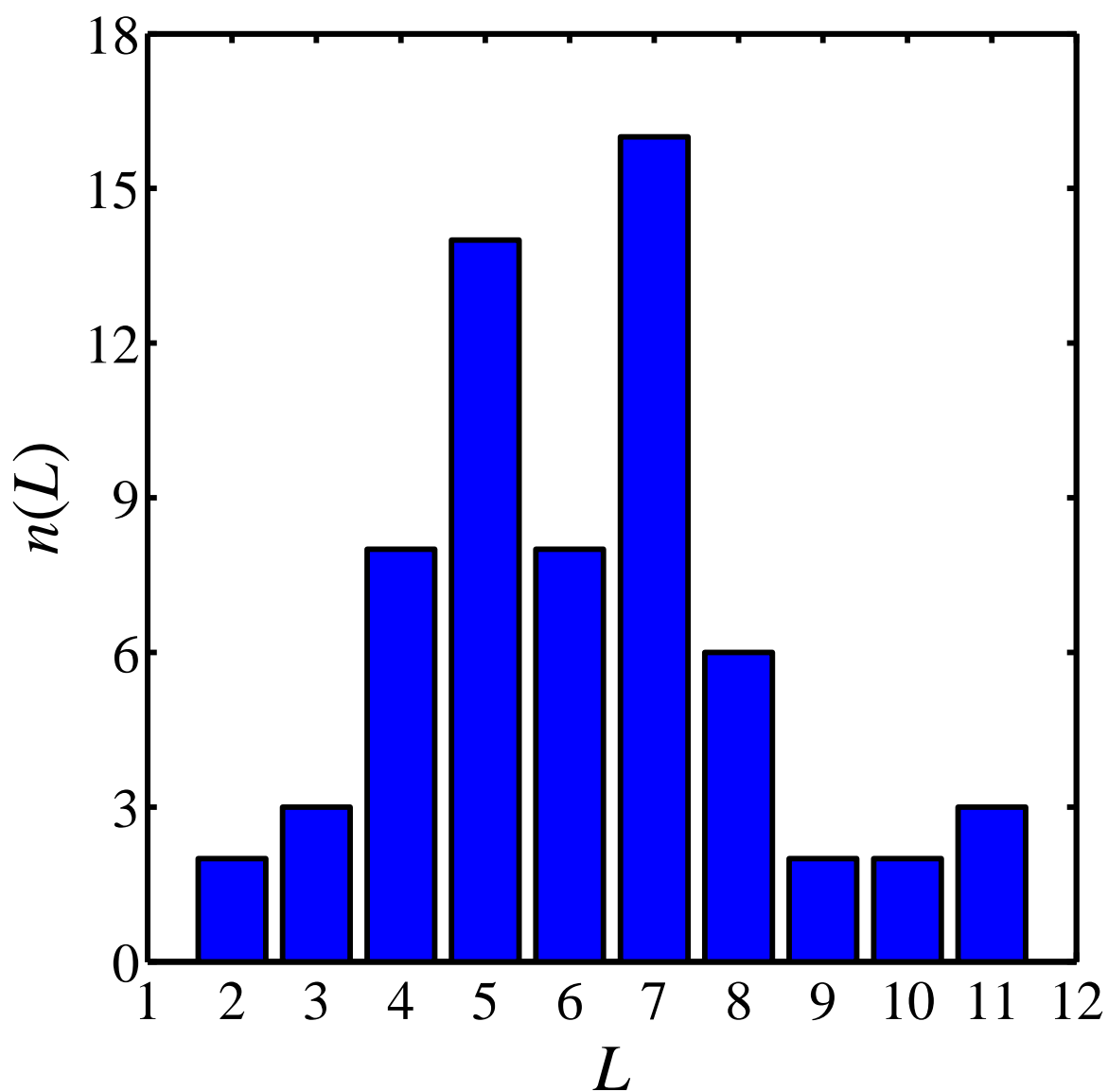
The problem:

Research by Stanley Milgram et al.  
Late 1960's.

- Target person worked in Boston as a stockbroker.
- $\simeq$  100 random senders from Boston.
- $\simeq$  100 senders from stockbrokers in Omaha, Nebraska.
- $\simeq$  100 random senders from Omaha, Nebraska.
- 20% of senders reached target.
- average chain length  $\simeq$  6.5.

The problem:

Travers and Milgram (Sociometry, 1969):  
“An Experimental Study of the Small World Problem.”



The problem:

Two significant features characterize a small-world network:

1. Short paths exist,

and

2. People are good at finding them.

Previous work—short paths:

Connected **random networks**  
have short average path lengths:

$$\langle x_{ij} \rangle \sim \log(N)$$

$N$  = population size,

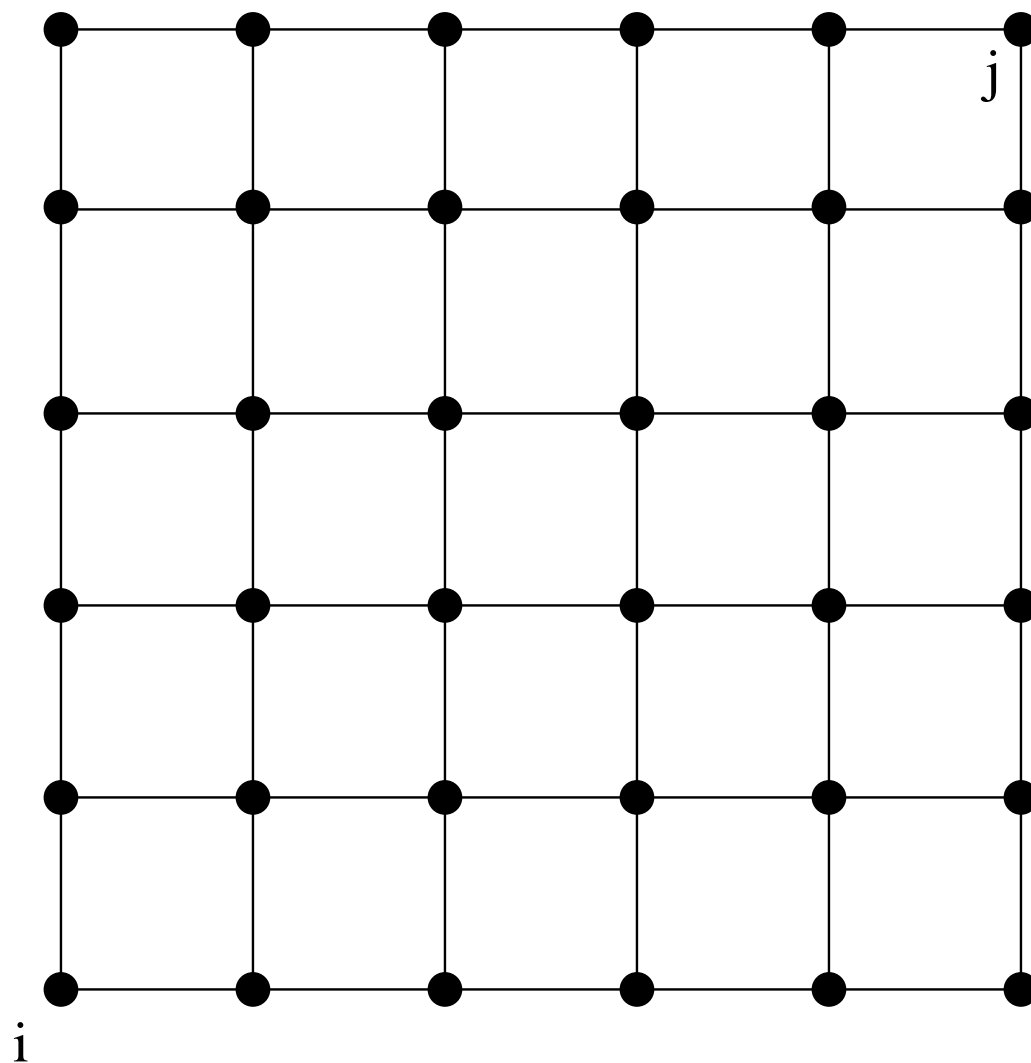
$x_{ij}$  = distance between nodes  $i$  and  $j$ .

**But: social networks aren't random.**



Previous work—short paths:

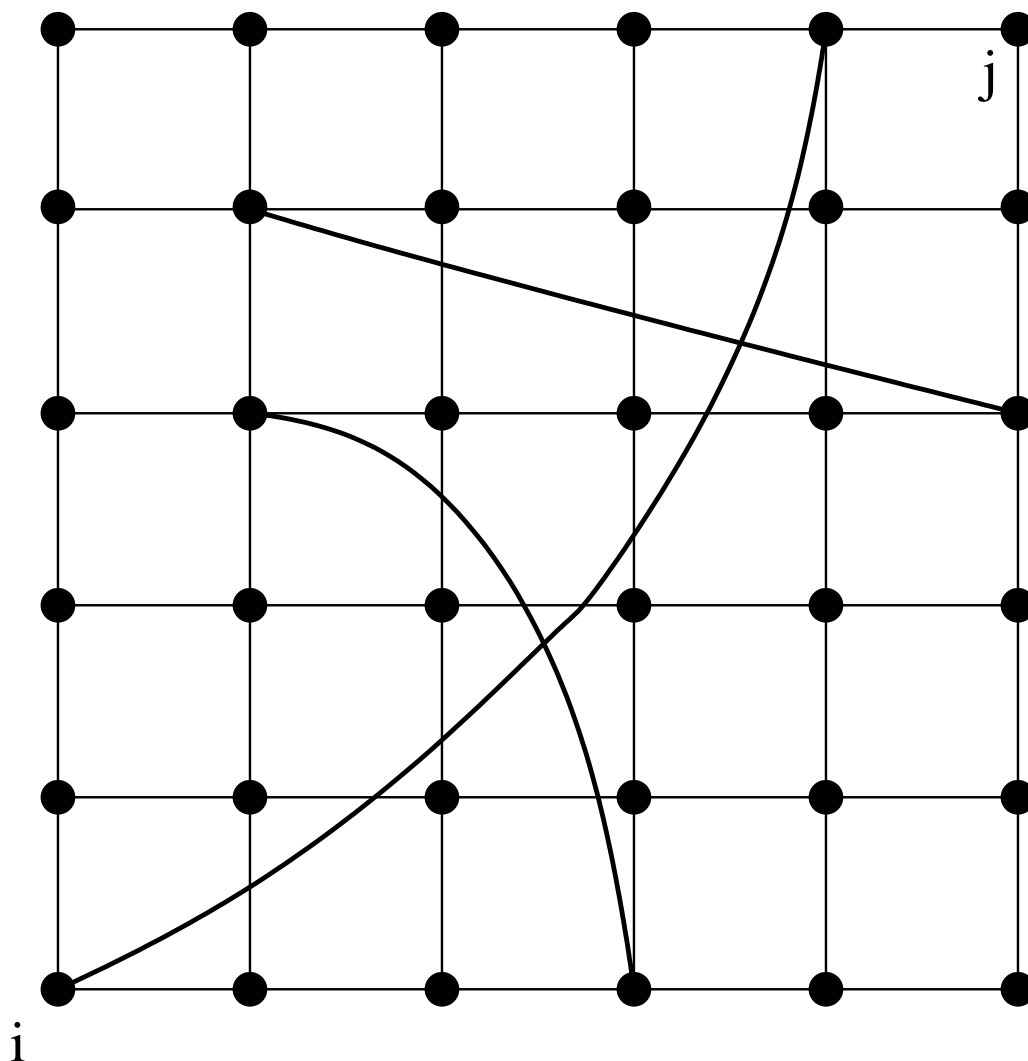
Non-randomness gives clustering:



But now  $x_{ij} = 10 \rightarrow$  too many long paths.

Previous work—short paths:

Resolution uses randomness and regularity:



Now have  $x_{ij} = 2$ .

$\langle x_{ij} \rangle$  goes down, clustering remains.

Previous work—short paths:

Introduced by  
Watts and Strogatz (Nature, 1998),  
“Collective dynamics of ‘small-world’  
networks.”

Small-world networks found everywhere:

- neural network of *C. elegans*,
- semantic networks of languages,
- actor collaboration graph,
- food webs.

**Very weak requirement:**

Regular underlying network structure

+

random **short cuts**.

Previous work—finding short paths:

But are these short cuts findable?

No.

Nodes cannot find each other quickly with any local search method.

Previous work—finding short paths:

What can a local search method use?

How to find things without a map?

Need some measure of distance between friends and the target.

Some possible knowledge:

1. Target's identity,
2. Friends' identities,
3. Friends' popularity,
4. Where message has been.

(Some of above may be partial).

Previous work—finding short paths:

Jon Kleinberg (Nature, 2000),  
“Navigation in a small world.”

Allowed to vary:

1. local search algorithm,  
and
2. network structure.

Previous work—finding short paths:

Network:

1. start with regular  $d$ -dimensional cubic lattice.
2. add local links so nodes know all nodes within a distance  $q$ .
3. add  $m$  short cuts per node between nodes  $i$  and  $j$  with probability

$$p_{ij} \propto x_{ij}^{-\alpha}.$$

$\alpha = 0$ : random connections.

$\alpha$  large: reinforce local connections.

$\alpha = d$ : same number of connections at all scales.

Previous work—finding short paths:

Theoretical optimal search:

1. “Greedy” algorithm.
2.  $\alpha = d$ .

Search time grows like  $\log^2(N)$ .

For  $\alpha \neq d$ , polynomial factor  $N^\beta$  appears.

But: social networks aren't lattices plus links.

Previous work—finding short paths:

If networks have **hubs** can also search well.

Adamic et al. (Phys. Rev. E, 2001), “Search in Power-Law Networks.”

$$P(k_i) \propto k_i^{-\gamma}$$

where  $k$  = degree of node  $i$  (number of friends).

Basic idea: get to hubs first  
(airline networks).

**But: hubs in social networks are limited.**

The problem:

If there are no hubs and no underlying lattice, how can search be efficient?

Which friend is closest to the target?

What does closest mean?

How to measure 'social distance' accurately?

The model:

One solution: incorporate **identity**.

**Identity** is formed from attributes such as:

1. Geographic location,
2. Type of employment,
3. Religious beliefs,
4. Recreational activities.

**Groups** are formed by people with at least one similar attribute.

The model:

Six propositions about social networks:

**Proposition 1:** Individuals have identities and belong to various groups that reflect these identities.

**Proposition 2:** Individuals break down the world into a hierarchy of categories.

The model:

A Geographic example: The United States.

Level 1: The country.

Level 3: Regions: South, North East, Midwest, West coast, South West, Alaska.

Level 4: States within regions  
(New York, Connecticut, Massachusetts, . . . ).

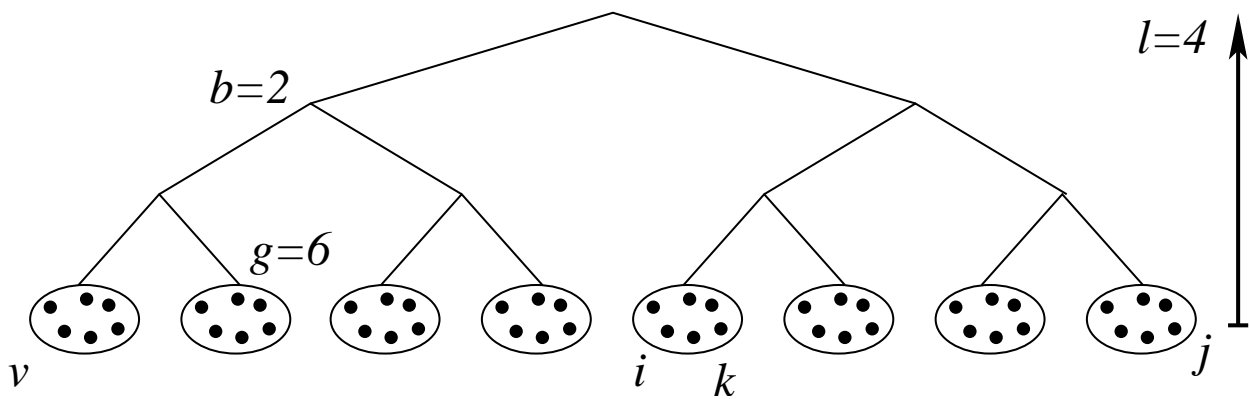
Level 5: Cities/areas within States  
(New York city, Boston, the Berkshires).

Level 6: Suburbs/towns/smaller cities  
(Brooklyn, Cambridge).

Level 7: Neighborhoods  
(the Village, Harvard Square).

The model:

Distance between two individuals  $x_{ij}$  is the lowest common ancestor in hierarchy.



$$x_{ij} = 3, x_{ik} = 1, x_{iv} = 4.$$

(Define distance between two individuals in same group as 1.)

The model:

**Proposition 3:** Individuals are more likely to know each other the closer they are within a hierarchy.

Construct  $z$  connections for each node using

$$p_{ij} = c \exp\{-\alpha x_{ij}\}.$$

$\alpha = 0$ : random connections.

$\alpha$  large: local connections.

The model:

**Proposition 4:** Each attribute of identity corresponds to a separate hierarchical classification of individuals.

Identity vector  $\vec{v}_i$ :

$v_i^h$  is position of node  $i$  in hierarchy  $h$ .

Basic example: two people live near each other but have different jobs.

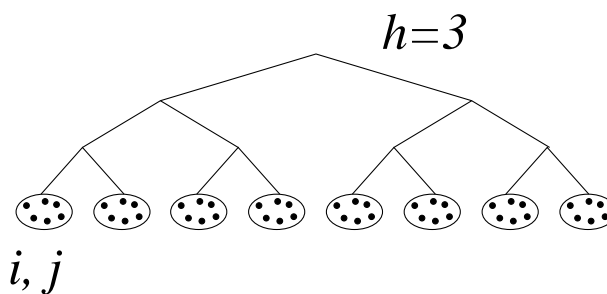
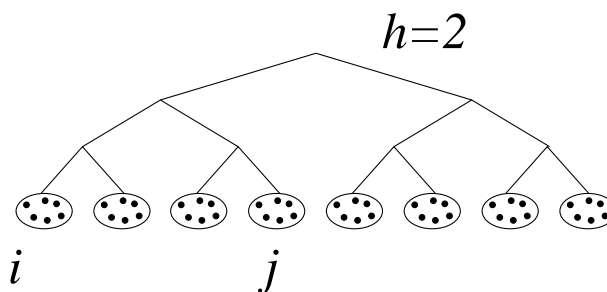
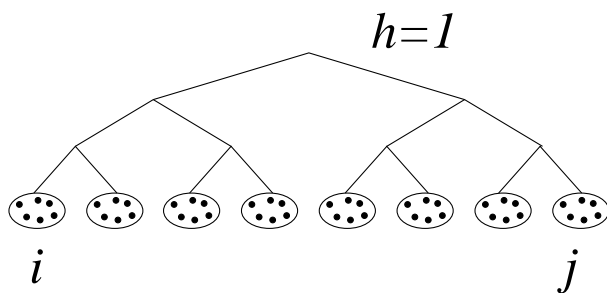
Now have  $h = 1, \dots, H$  hierarchies.

Individuals divide friends up between hierarchies.

Still connect with  $p_{ij}$ .

Model assumes no correlation between hierarchies.

The model:



$$\vec{v}_i = [1 \ 1 \ 1]^T, \quad \vec{v}_j = [8 \ 4 \ 1]^T.$$

$$x_{ij}^1 = 4, \quad x_{ij}^2 = 3, \quad x_{ij}^3 = 1.$$

The model:

**Proposition 5:** “Social distance” is the minimum distance between two nodes in all hierarchies.

$$y_{ij} = \min_h x_{ij}^h.$$

To be close means only one attribute has to be the same.

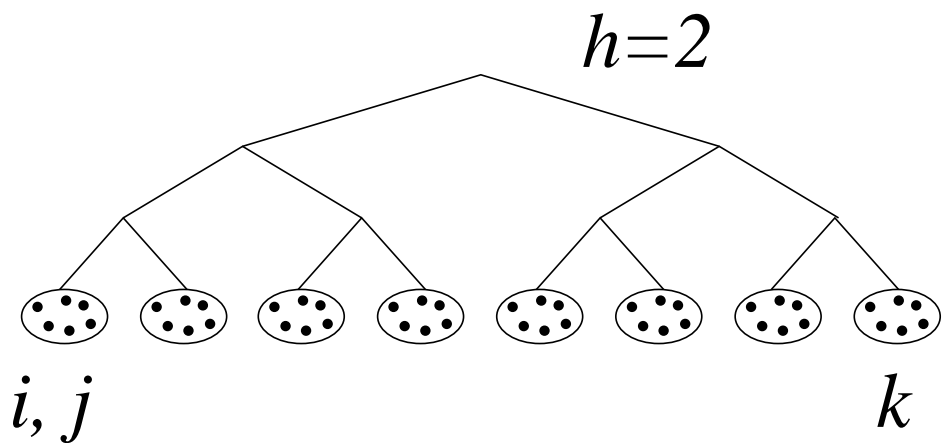
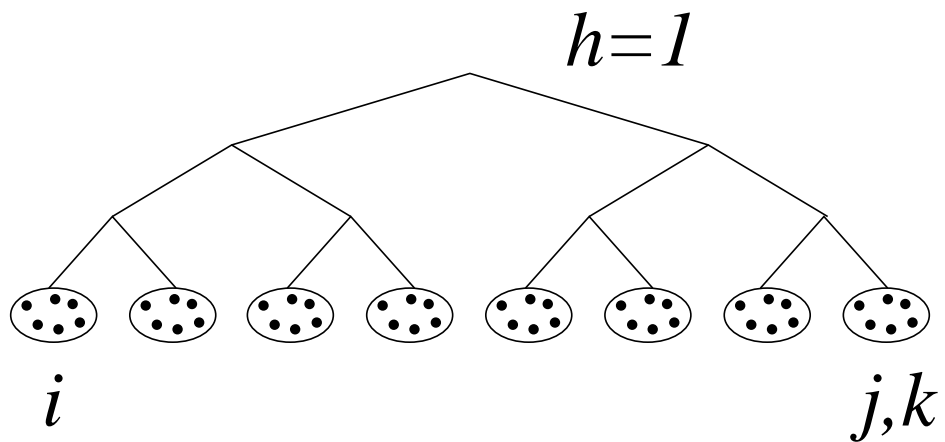
Previous slide:

$$x_{ij}^1 = 4, \quad x_{ij}^2 = 3, \quad x_{ij}^3 = 1.$$

$$\Rightarrow y_{ij} = 1.$$

The model:

Triangle inequality doesn't hold:



$$y_{ik} = 4 > y_{ij} + y_{jk} = 1 + 1 = 2.$$

The model:

**Proposition 6:** Individuals know the identity vectors of

1. themselves,
2. their friends,

and

3. the target.

Individuals can estimate the social distance between their friends and the target.

Use a greedy algorithm.

The model:

Define  $q$  as probability of an arbitrary message chain reaching a target.

Definition of a [searchable network](#):

Any network for which

$$q \geq r$$

for a desired  $r$ .

The model:

If message chains fail at each node with probability  $p$ , require

$$q = \langle (1 - p)^L \rangle \geq r.$$

where  $L =$  length of message chain.

Approximation:

$$\langle L \rangle \leq \ln r / \ln (1 - p).$$

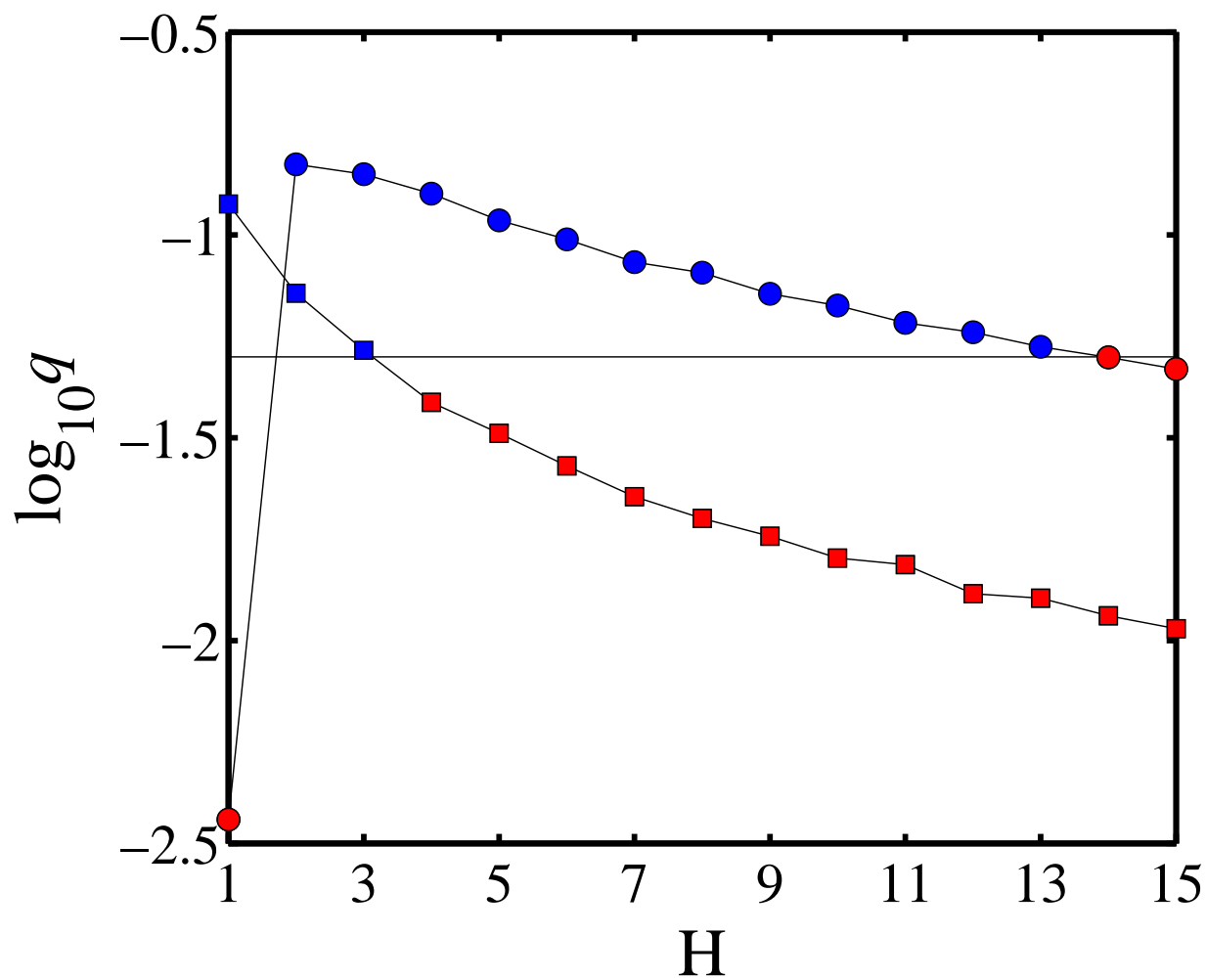
For  $r = 0.05$  and  $p = 0.25$ ,

$$\langle L \rangle \leq 10.4$$

independent of  $N$ .

The model-results:

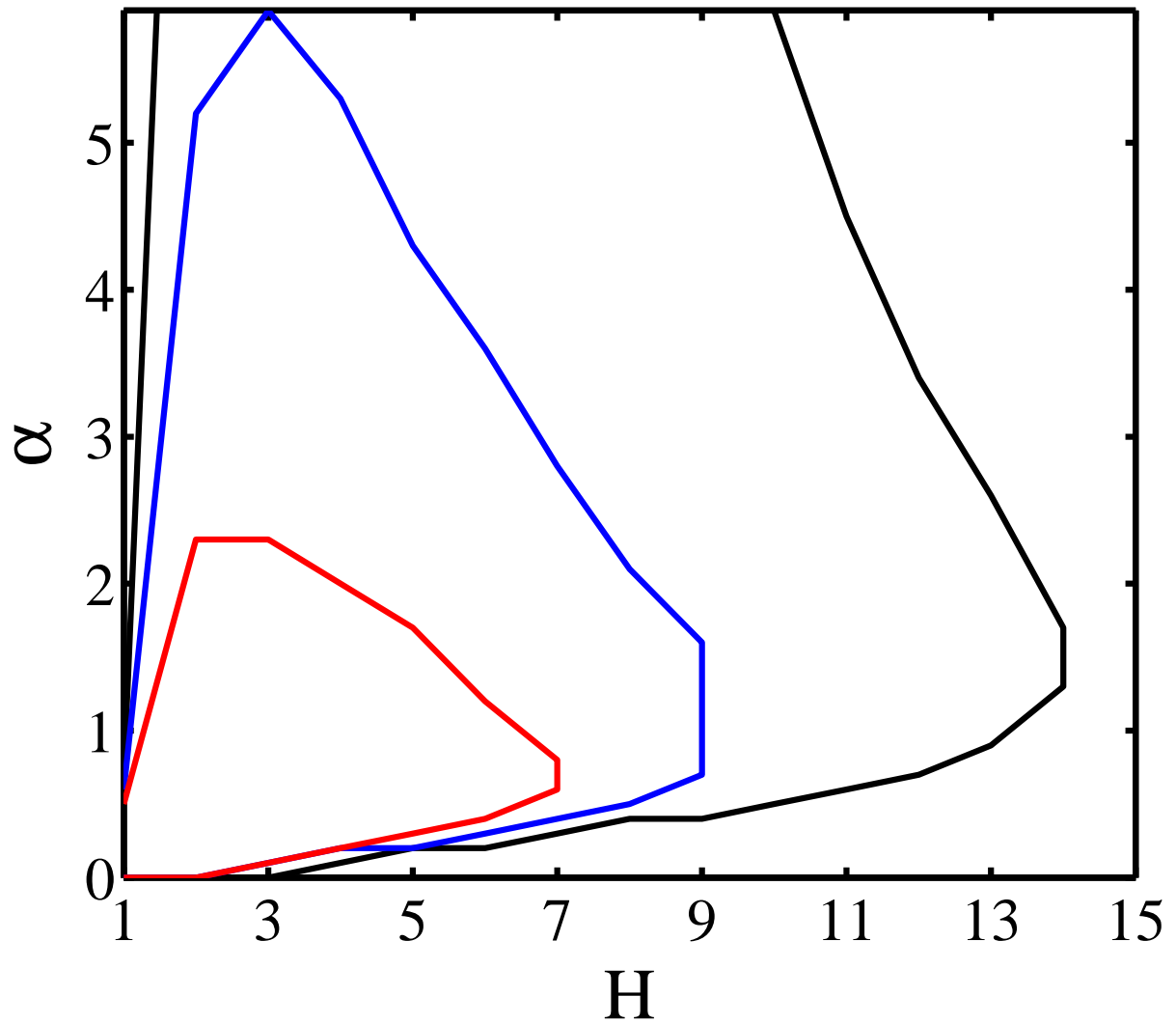
$\alpha = 0$  versus  $\alpha = 2$  for  $N=102400$ :



$q \geq r$

$q < r$

The model-results:



N=102400

p=0.25, r=0.05

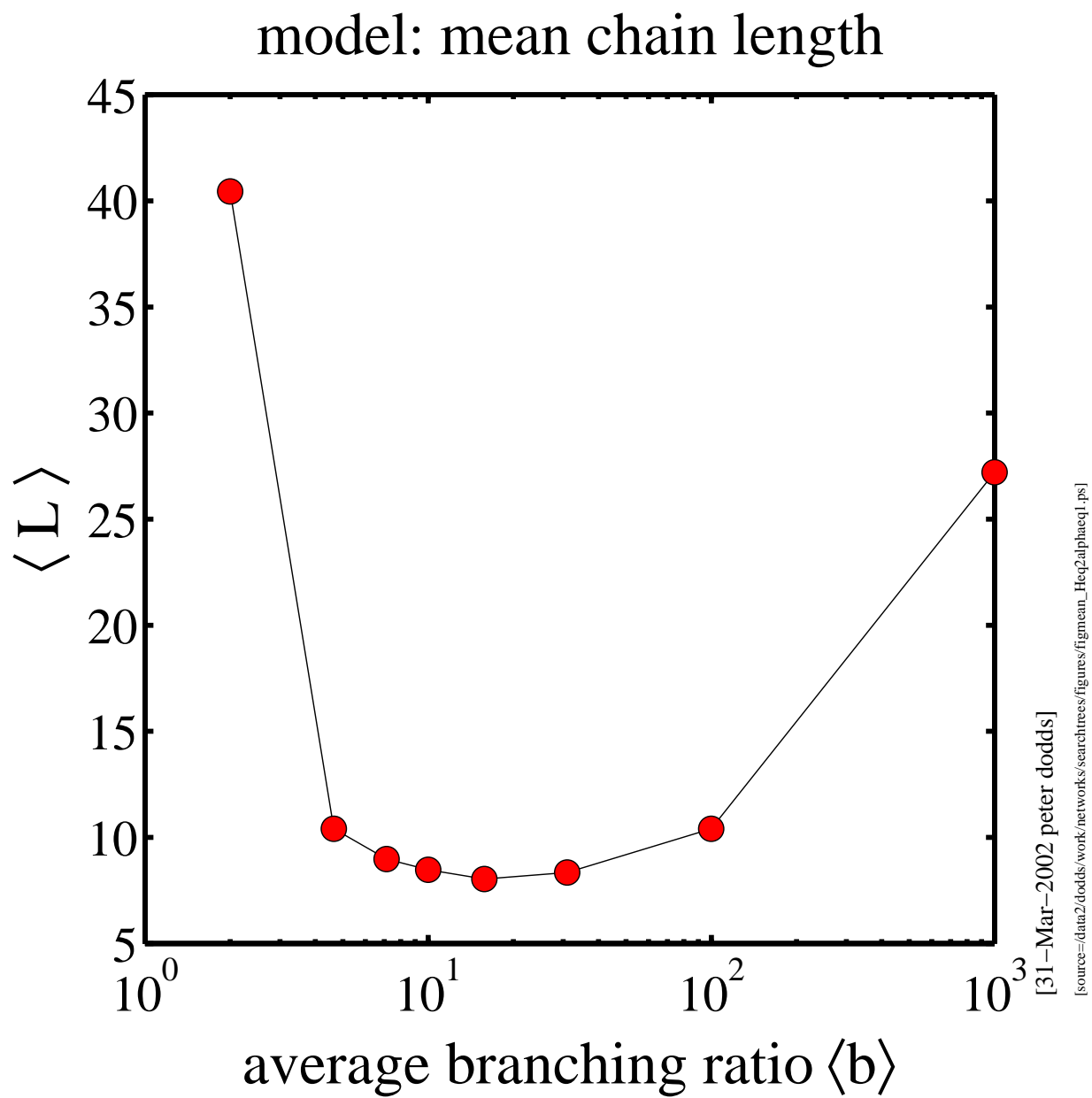
N=204800

N=409600

b=2, g=100, z=99

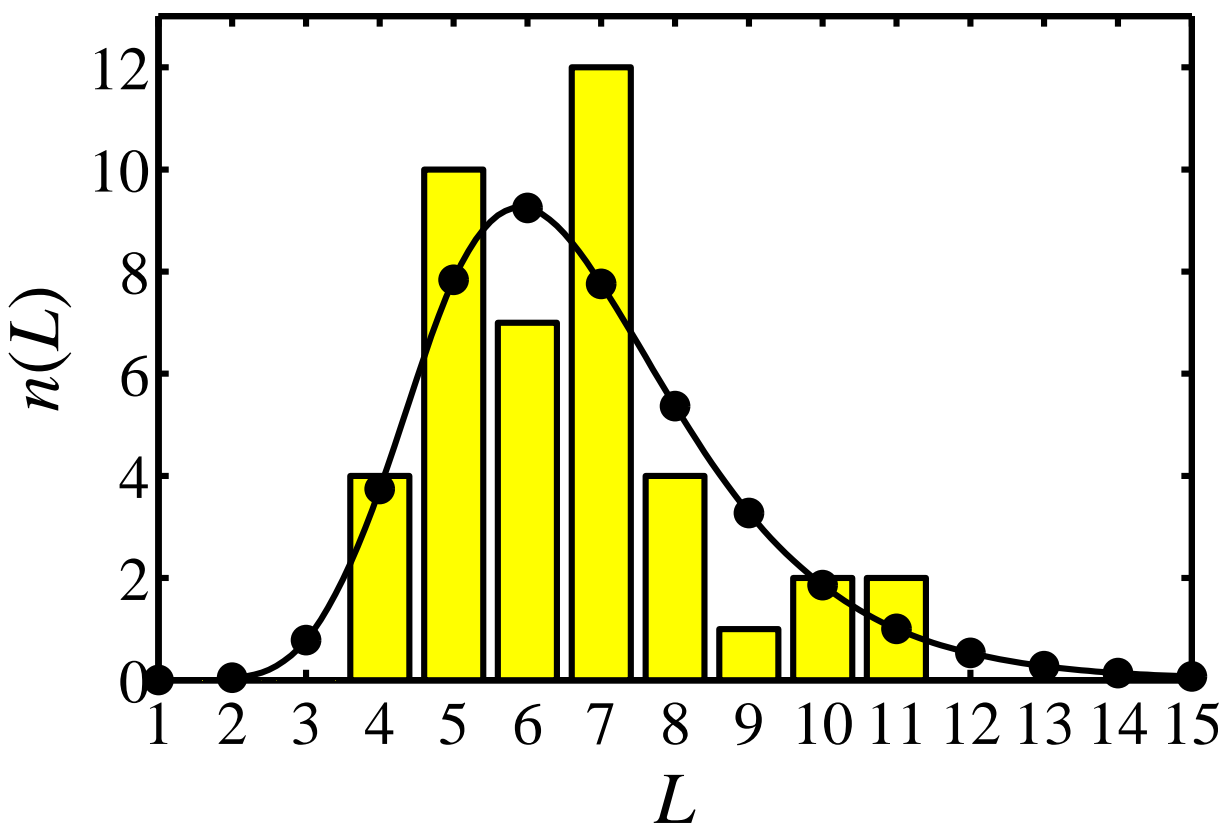
The model-results:

$N \simeq 10^8$ :



The model-results:

Milgram's Nebraska-Boston data:



Model parameters:

$$N = 10^8,$$

$$z = 300, g = 100,$$

$$b = 10,$$

$$\alpha = 1, H = 2.$$

$$\langle L_{\text{model}} \rangle \simeq 6.7$$

$$\langle L_{\text{data}} \rangle \simeq 6.5$$

Conclusions:

- Paths are findable if nodes understand how network is formed.
- Identity is crucial.
- Construction of peer-to-peer networks.
- Search in information databases.
- Sociology can help physics.