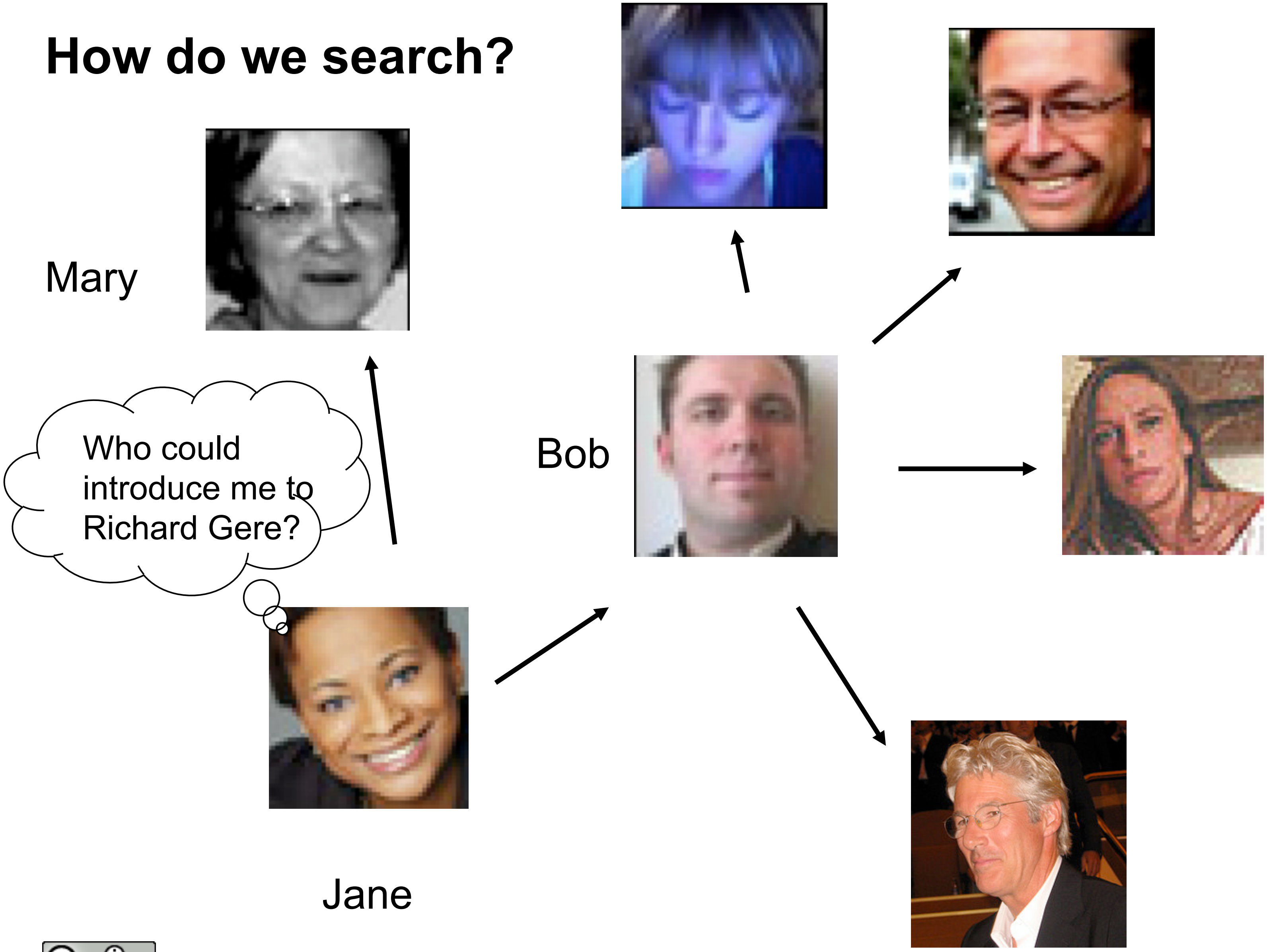
Search in structured networks

How do we search?



Mary

Who could introduce me to Richard Gere?

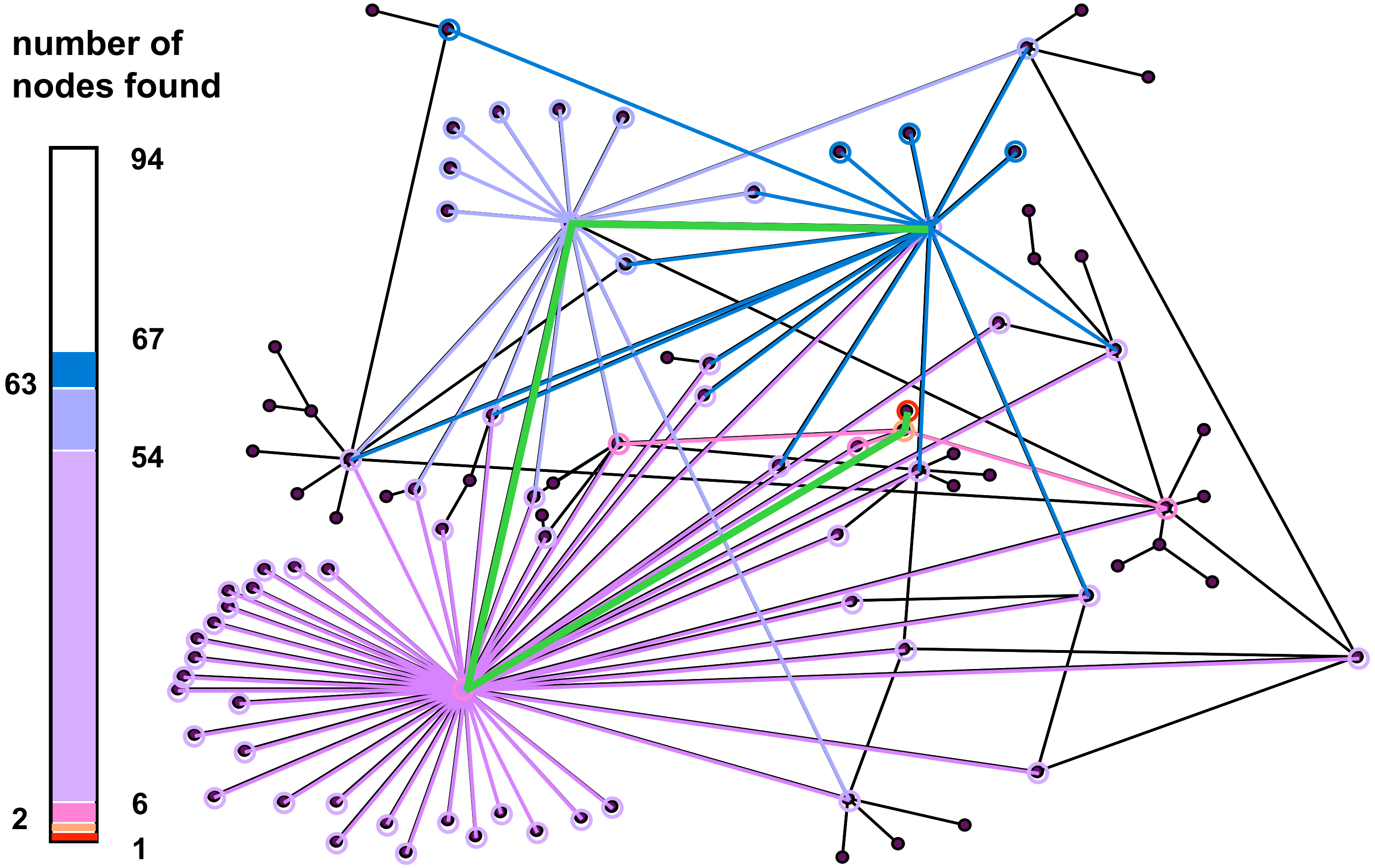
Bob

Jane

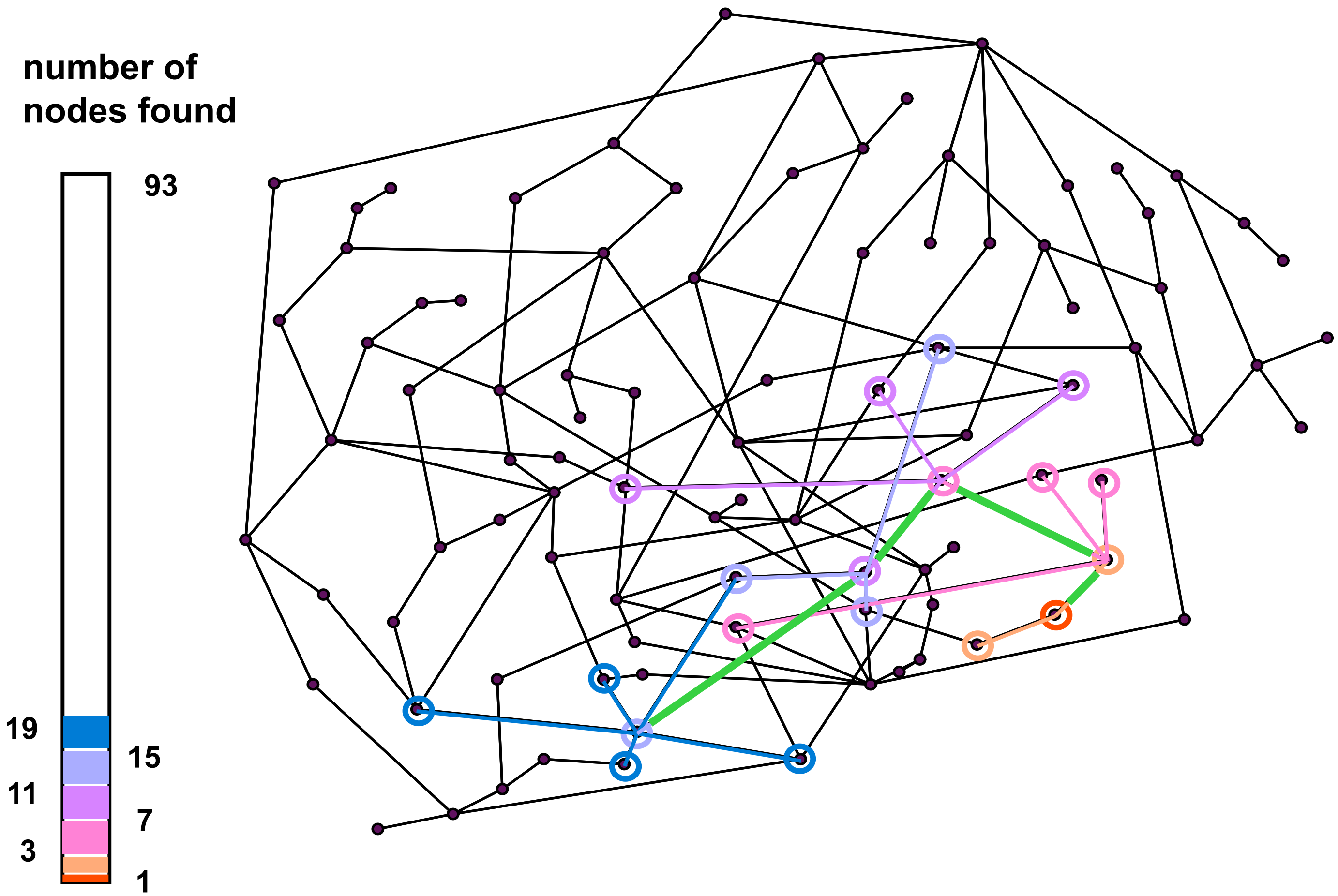
Richard Gere – spaceodissey, Flickr; http://creativecommons.org/licenses/by/2.0/deed.en

Friends collage – luc, Flickr; http://creativecommons.org/licenses/by/2.0/deed.en

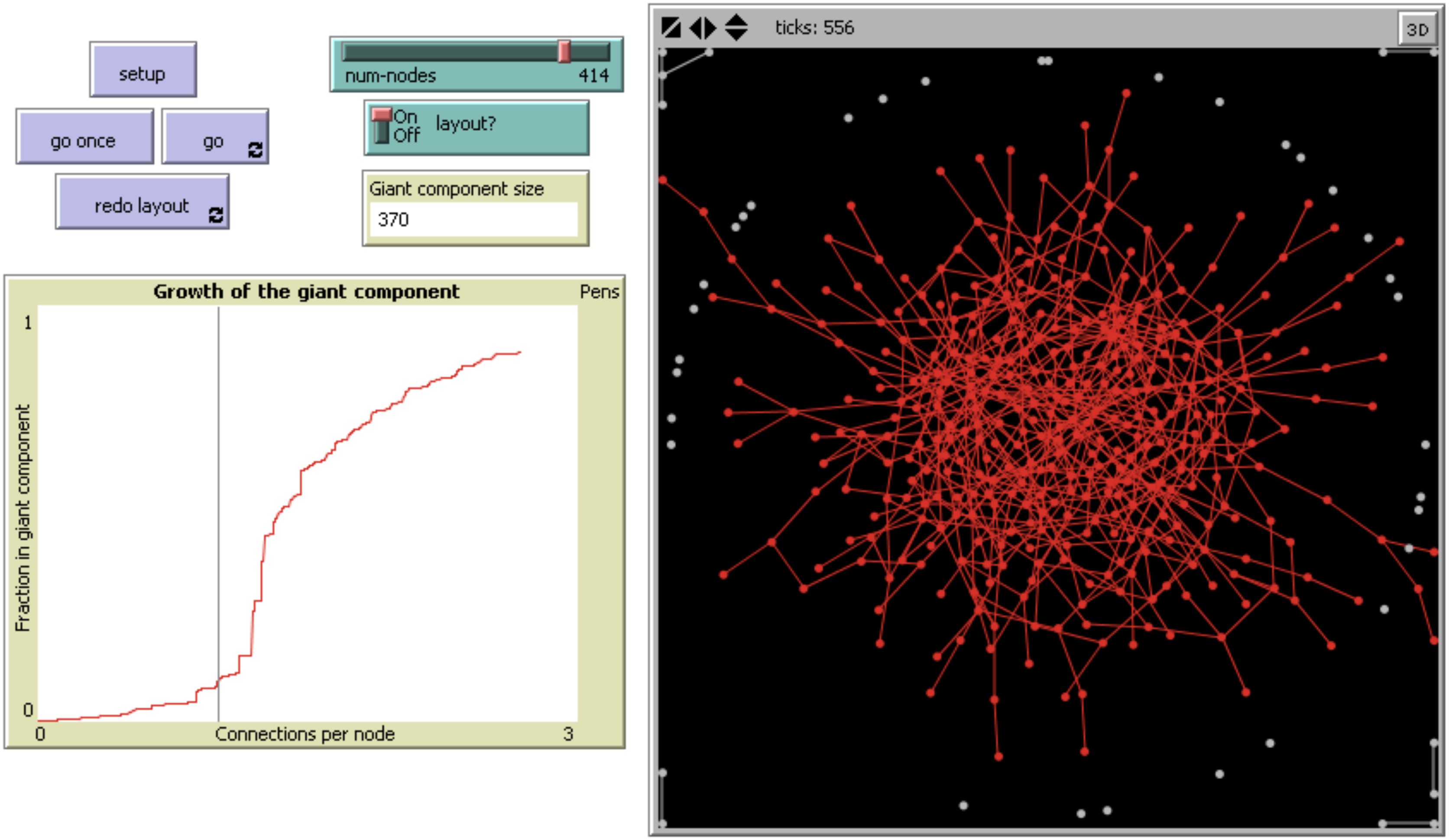
power-law graph



Poisson graph



How would you search for a node here?



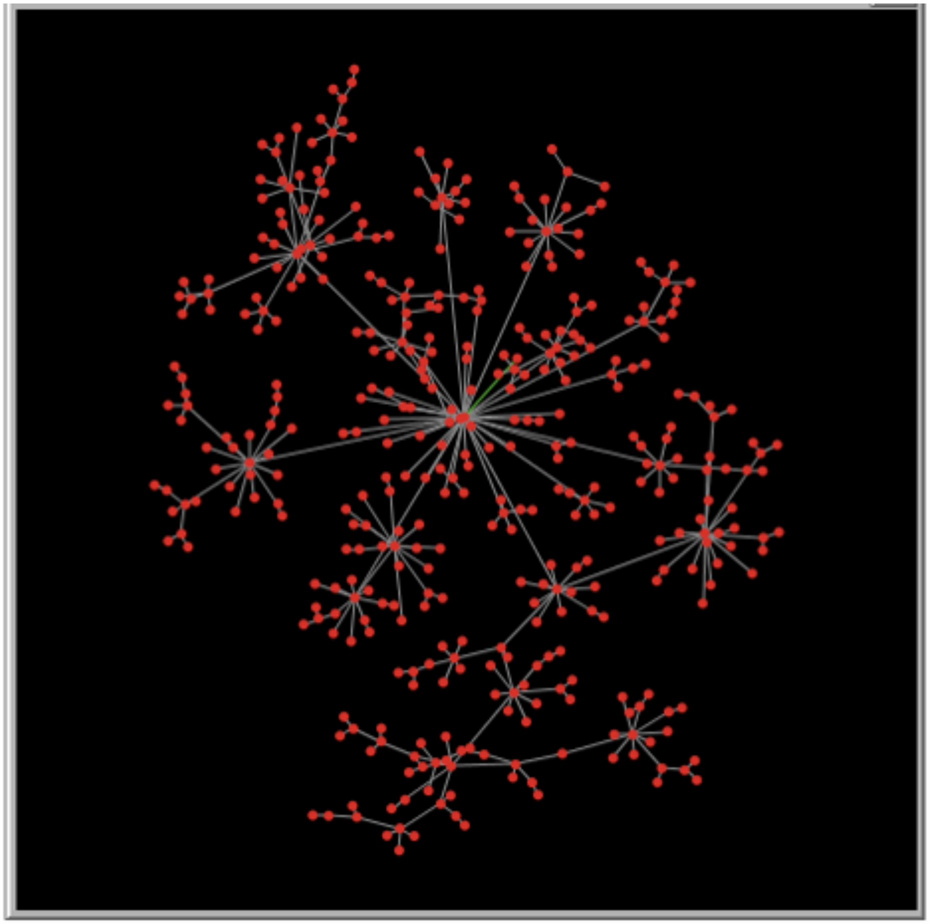
num-nodes 414

layout?

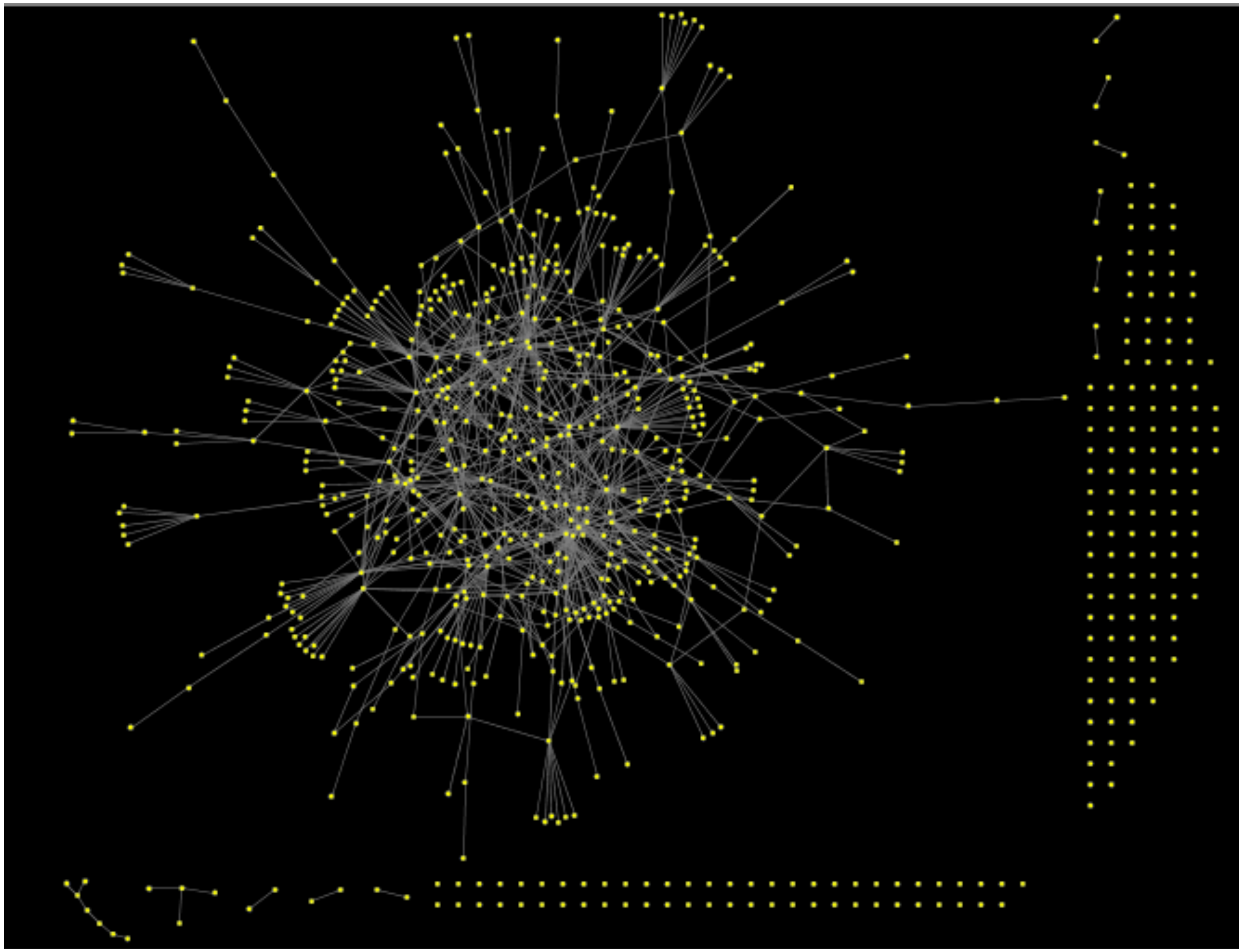
Giant component size 370

ticks: 556

What about here?

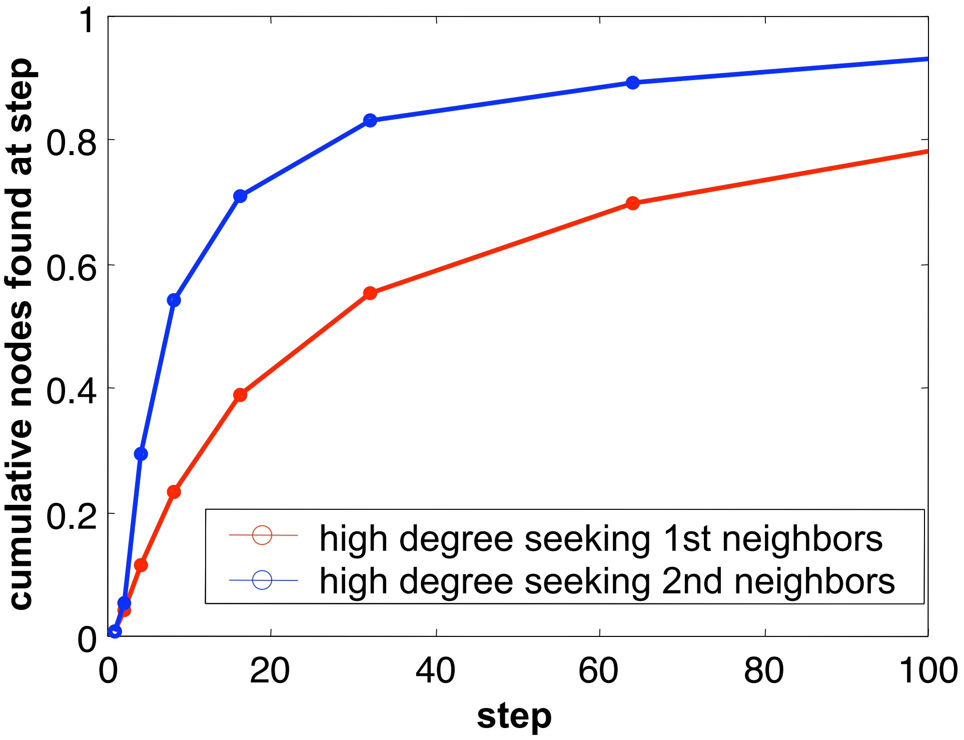


gnutella network fragment

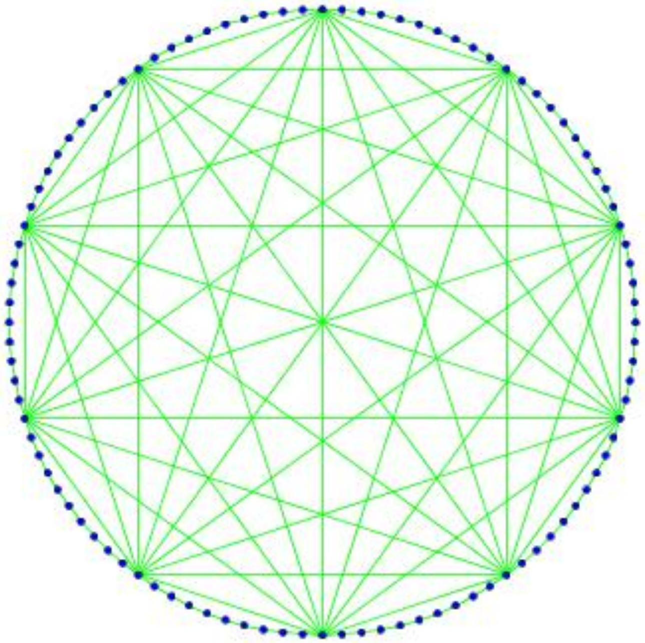


Gnutella network

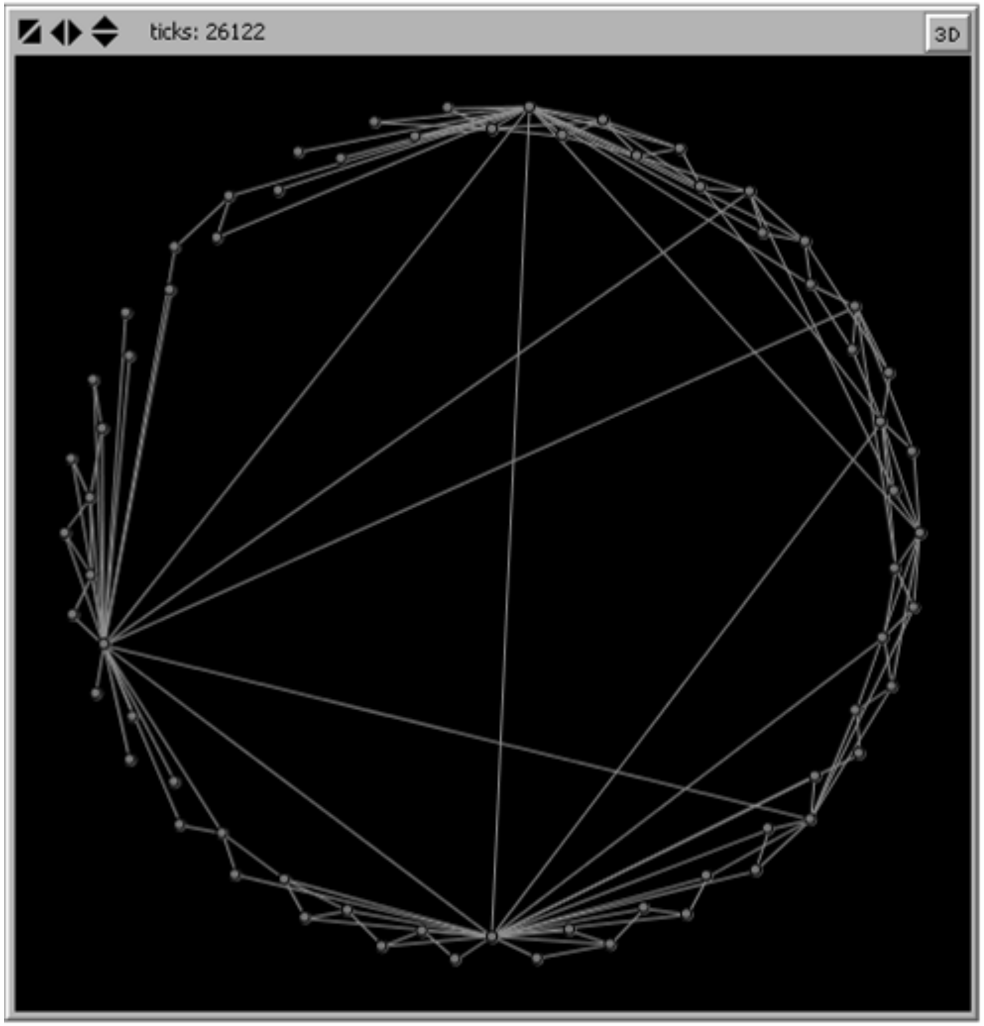
50% of the files in a 700 node network can be found in < 8 steps



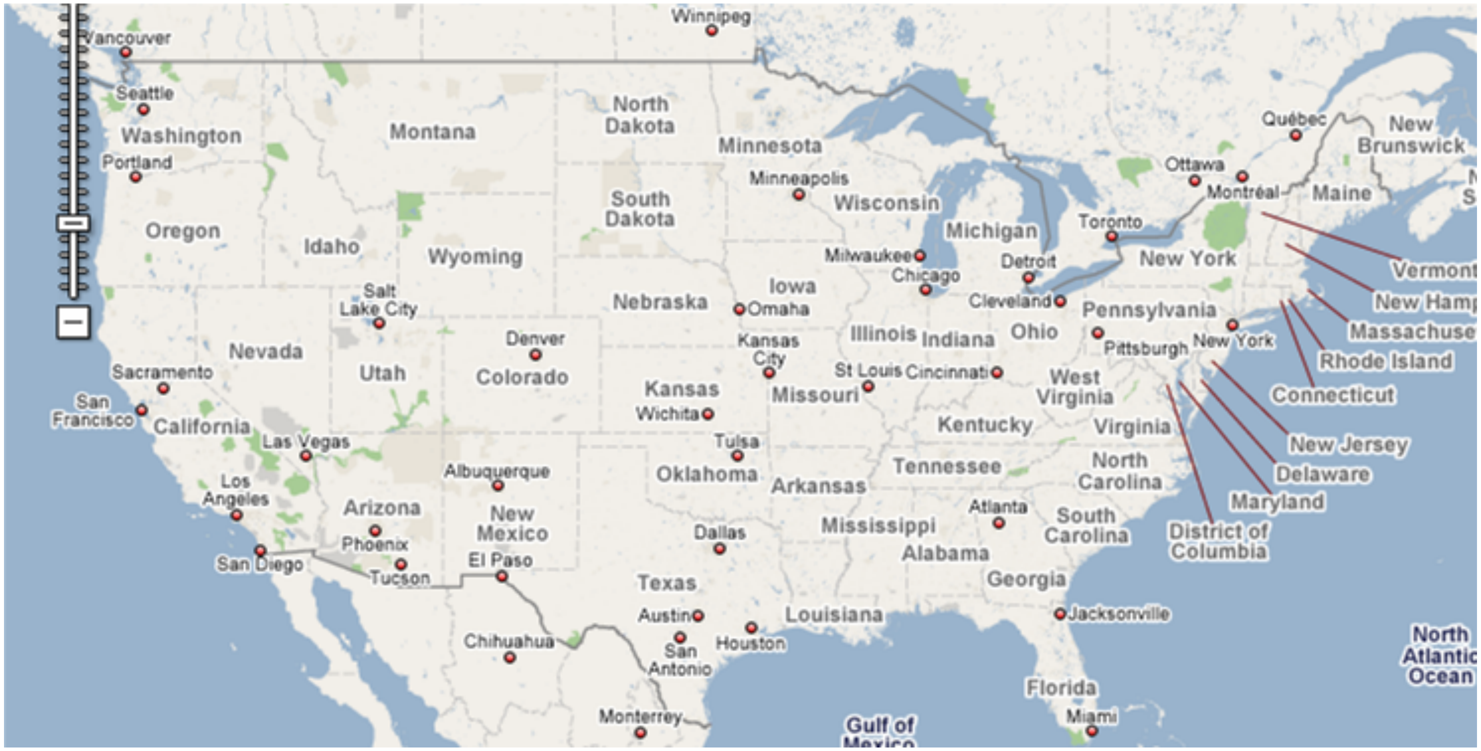
And here?



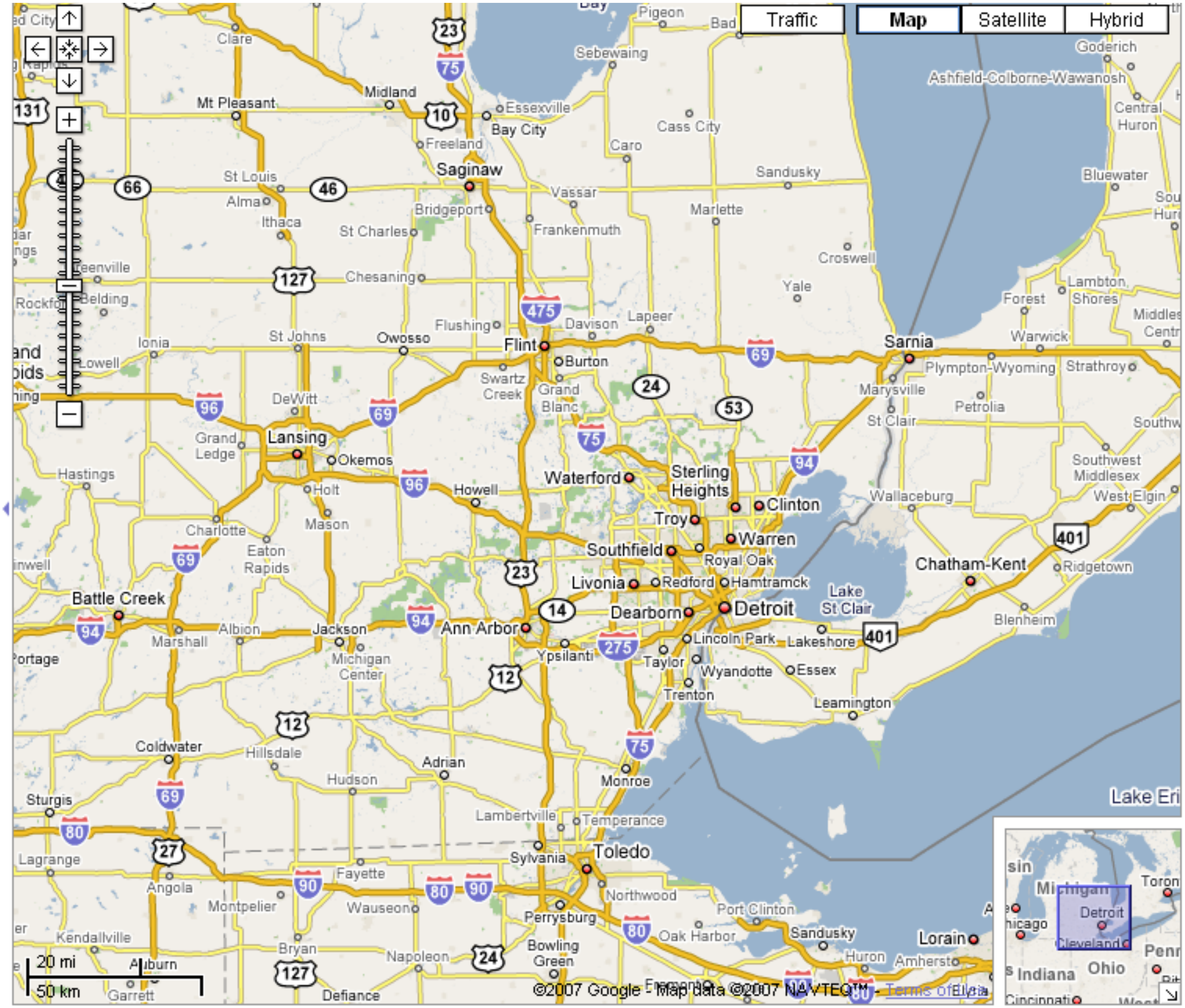
here?



here?

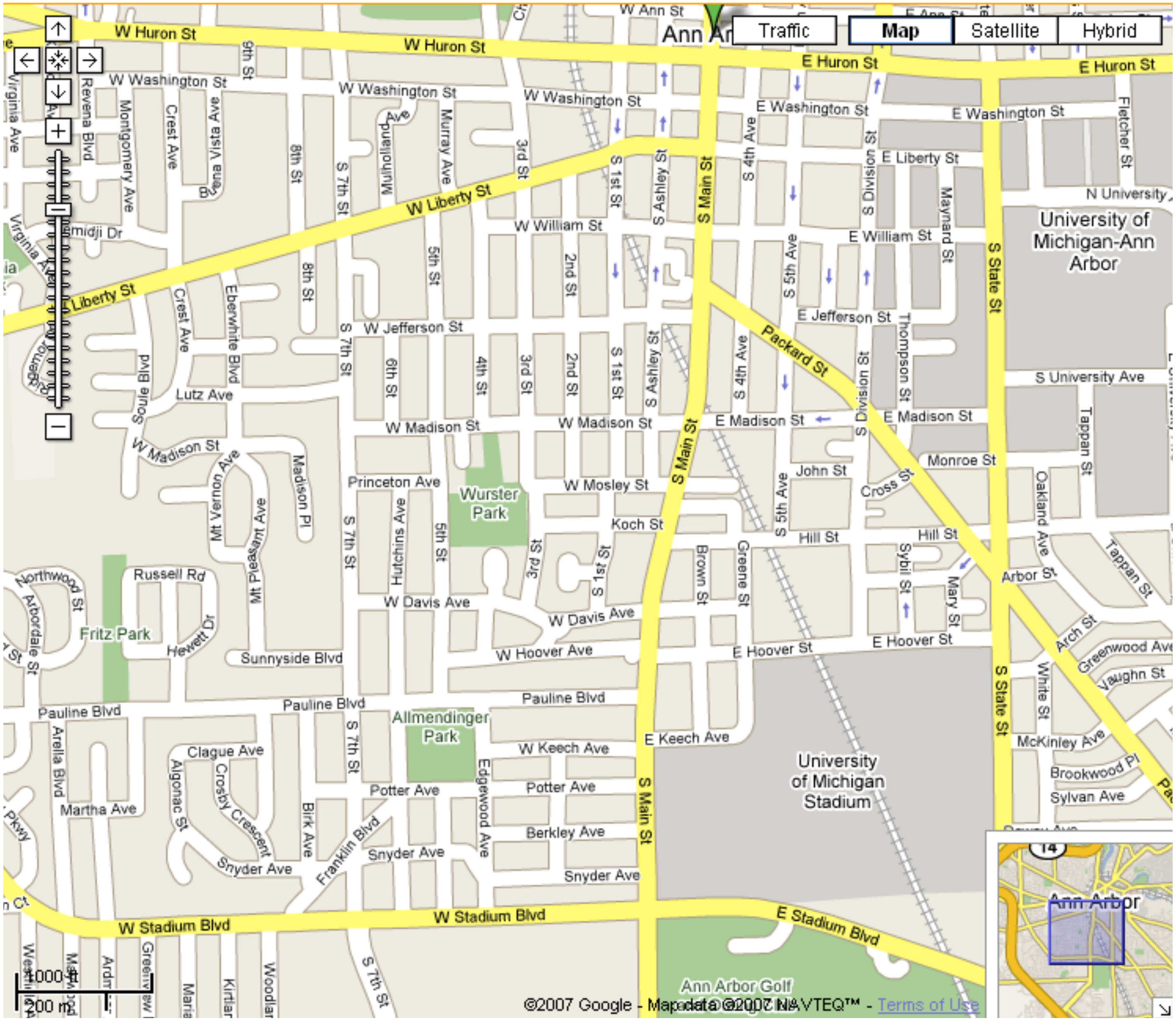


Source: http://maps.google.com



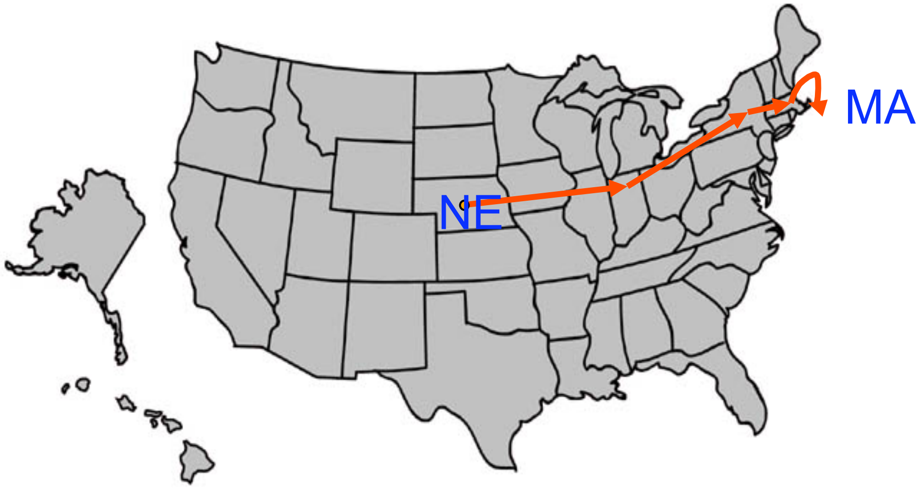
Source: http://maps.google.com

here?



Source: http://maps.google.com

Small world experiments review



Source: undetermined

Source: NASA, U.S. Government; http://visibleearth.nasa.gov/view_rec.php?id=2429

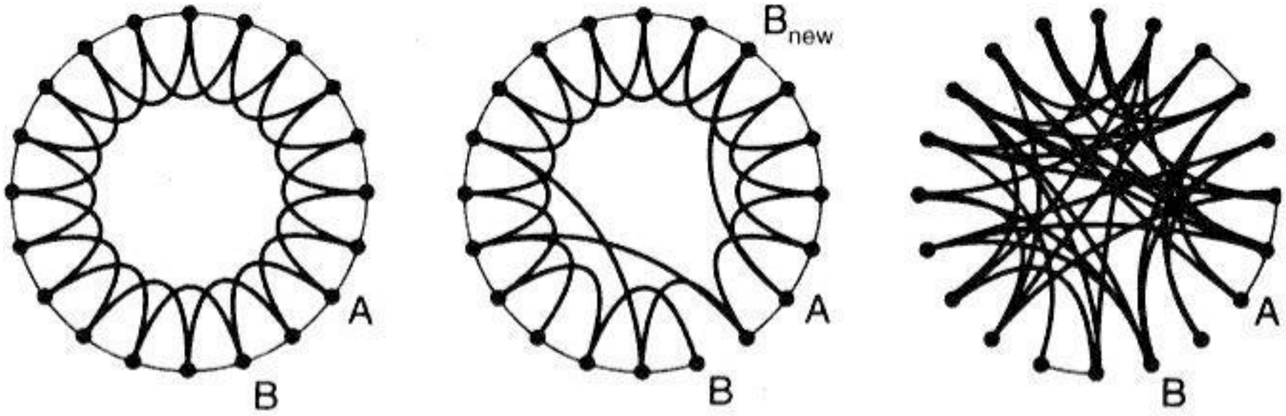
Milgram (1960's), Dodds, Muhamad, Watts (2003)

Given a target individual and a particular property, pass the message to a person you correspond with who is "closest" to the target.

Short chain lengths – six degrees of separation

Typical strategy – if far from target choose someone geographically closer, if close to target geographically, choose someone professionally closer

Is this the whole picture?



Why are small worlds navigable?

Source: Watts, D.J., Strogatz, S.H.(1998) Collective dynamics of 'small-world' networks. Nature 393:440-442.

How are people are able to find short paths?

How to choose among hundreds of acquaintances?

Strategy:
Simple greedy algorithm - each participant chooses
correspondent
who is closest to target with respect to the given property

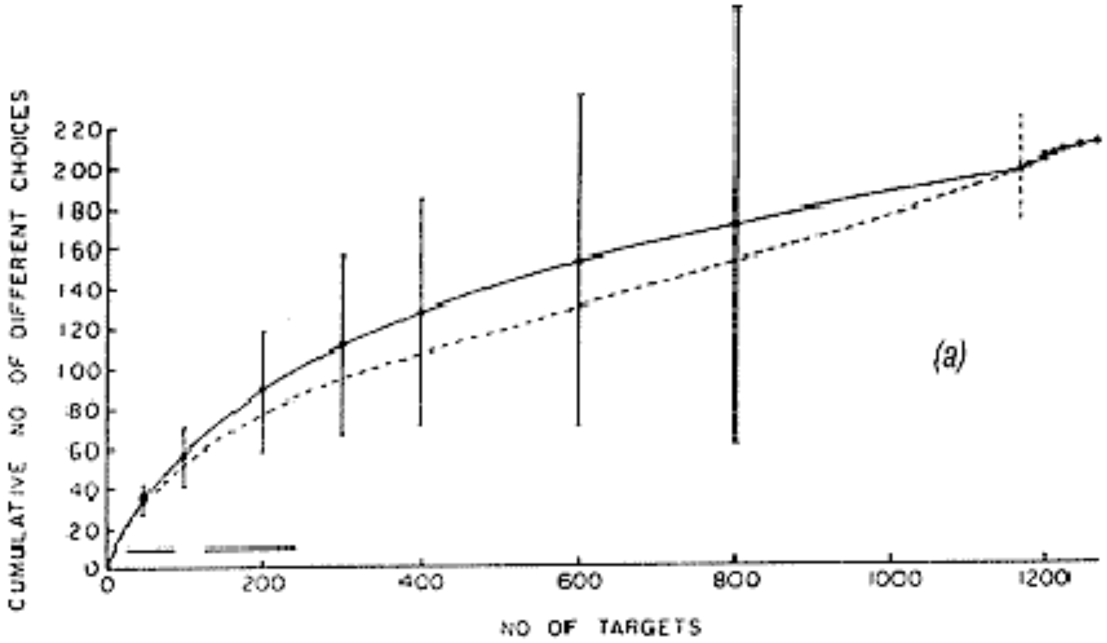
Models

geography
Kleinberg (2000)

hierarchical groups
Watts, Dodds, Newman (2001), Kleinberg(2001)

high degree nodes
Adamic, Puniyani, Lukose, Huberman (2001), Newman(2003)

Reverse small world experiment



- Killworth & Bernard (1978):
- Given hypothetical targets (name, occupation, location, hobbies, religion…) participants choose an acquaintance for each target
- Acquaintance chosen based on
- (most often) occupation, geography
- only 7% because they "know a lot of people"
- Simple greedy algorithm: most similar acquaintance
- two-step strategy rare

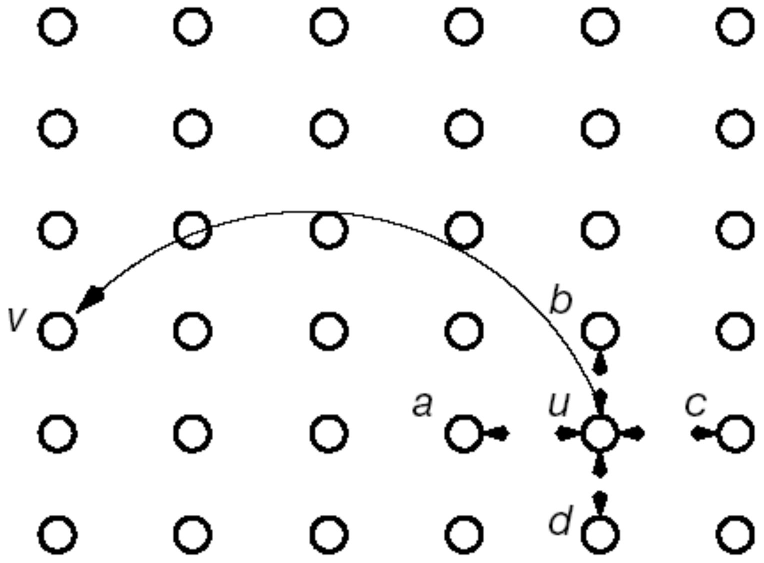
Source: 1978 Peter D. Killworth and H. Russell Bernard. The Reverse Small World Experiment Social Networks 1:159–92.

How many hops actually separate any two individuals in the world?

- Participants are not perfect in routing messages
- They use only local information
- **"The accuracy of small world chains in social networks"** Peter D. Killworth, Chris McCarty , H. Russell Bernard& Mark House:
  - Analyze 10920 shortest path connections between 105 members of an interviewing bureau,
  - together with the equivalent conceptual, or 'small world' routes, which use individuals' selections of intermediaries.
  - This permits the first study of the impact of accuracy within small world chains.
  - The mean small world path length (3.23) is 40% longer than the mean of the actual shortest paths (2.30)
  - Model suggests that people make a less than optimal small world choice more than half the time.

review: Spatial search

Kleinberg, 'The Small World Phenomenon, An Algorithmic Perspective'
Proc. 32nd ACM Symposium on Theory of Computing, 2000.
(Nature 2000)



"The geographic movement of the [message] from Nebraska to Massachusetts is striking. There is a progressive closing in on the target area as each new person is added to the chain"
S.Milgram 'The small world problem', Psychology Today 1,61,1967

nodes are placed on a lattice and connect to nearest neighbors

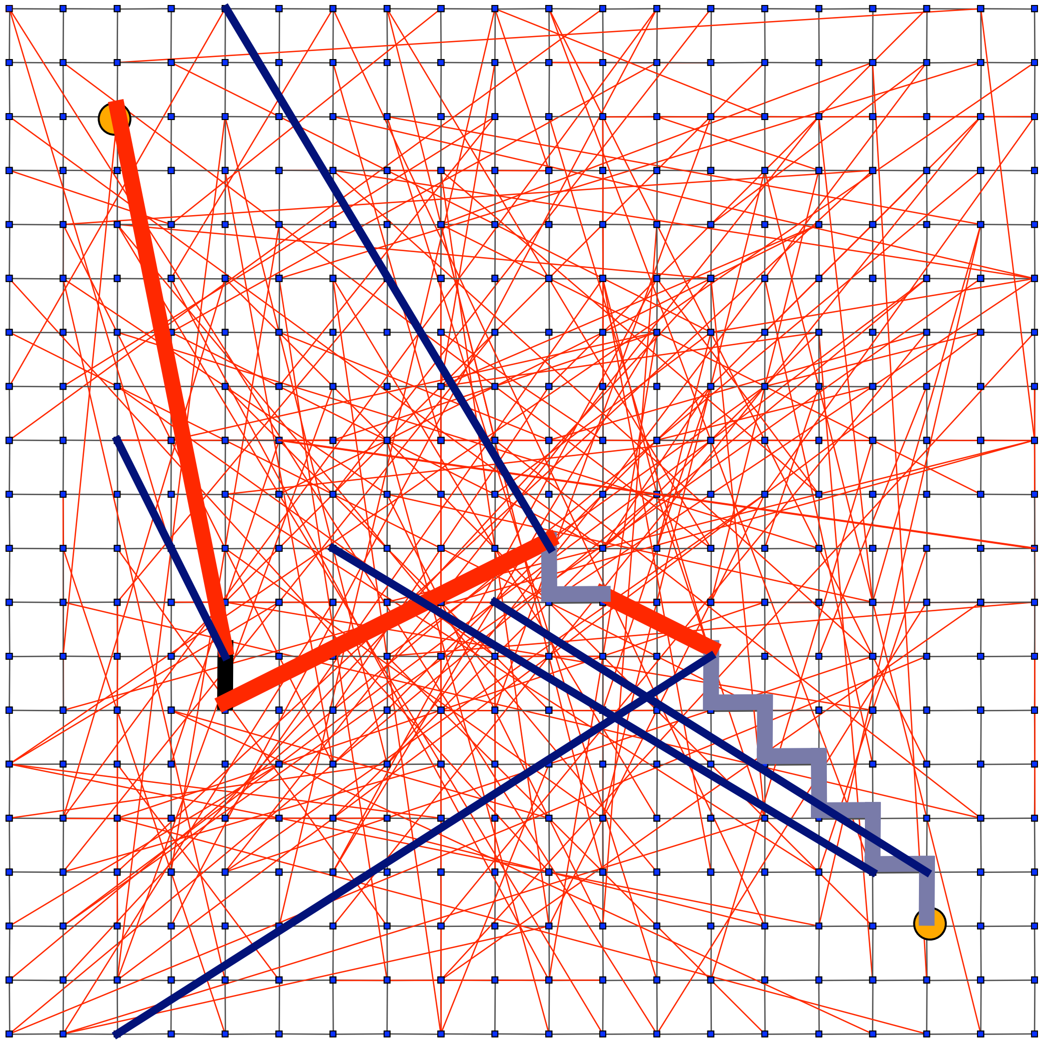
additional links placed with $p_{uv} \sim d_{uv}^{-r}$

no locality

When **r=0**, links are randomly distributed, ASP ~ **log(n)**, n size of grid

When **r=0**, any decentralized algorithm is at least $\mathbf{a_0 n^{2/3}}$

$p \sim p_0$

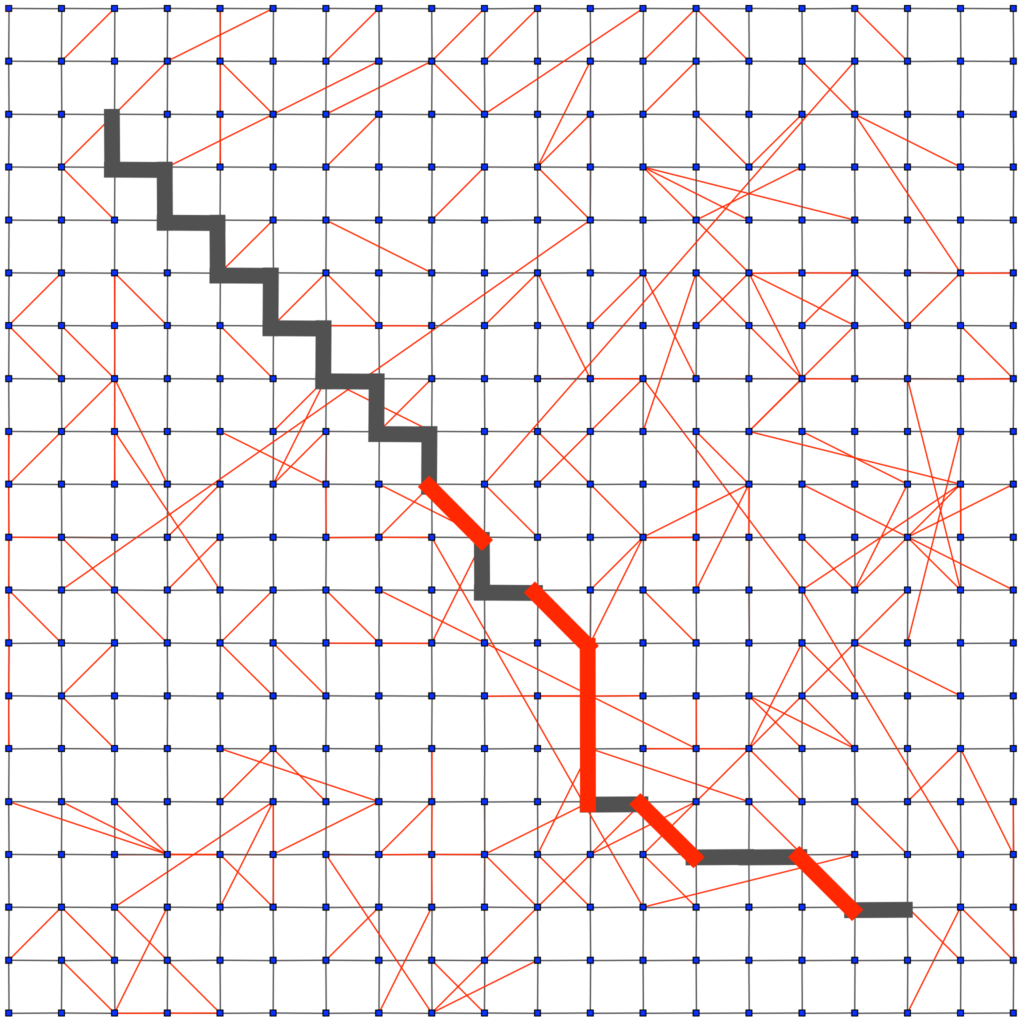


When **r<2**, expected time at least $\alpha_r n^{(2-r)/3}$

Overly localized links on a lattice

When **r>2** expected search time ~ $\mathbf{N^{(r-2)/(r-1)}}$

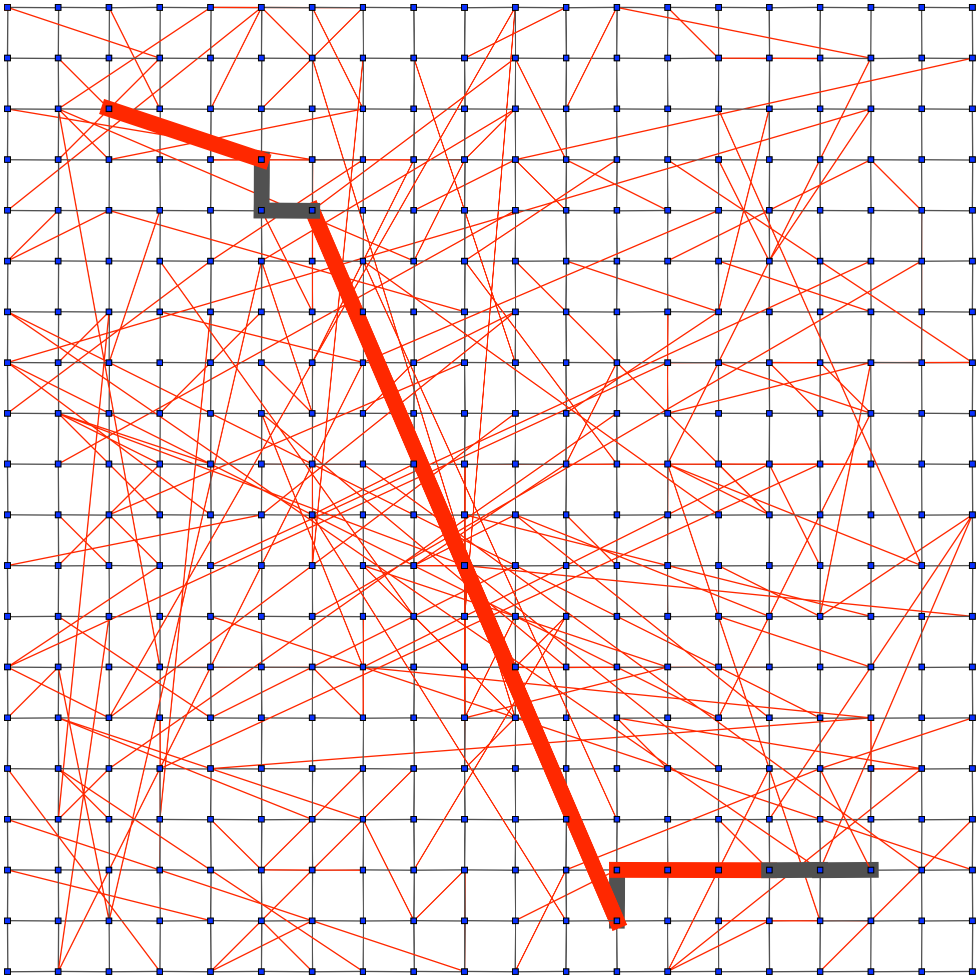
$$p \sim \frac{1}{d^4}$$



Links balanced between long and short range

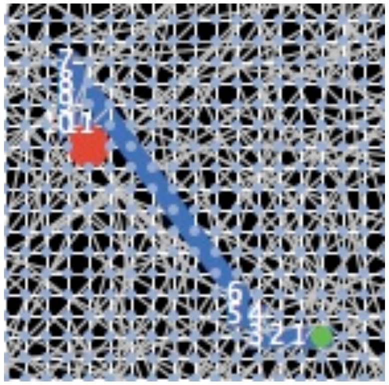
When **r=2,** expected time of a DA is at most **C (log N)^2**

$$p \sim \frac{1}{d^2}$$



demo

- how does the probability of long-range links affect search?



http://projects.si.umich.edu/netlearn/NetLogo4/SmallWorldSearch.html

Testing search models on social networks

advantage: have access to entire communication network
and to individual's attributes

Use a well defined network:
HP Labs email correspondence over 3.5 months

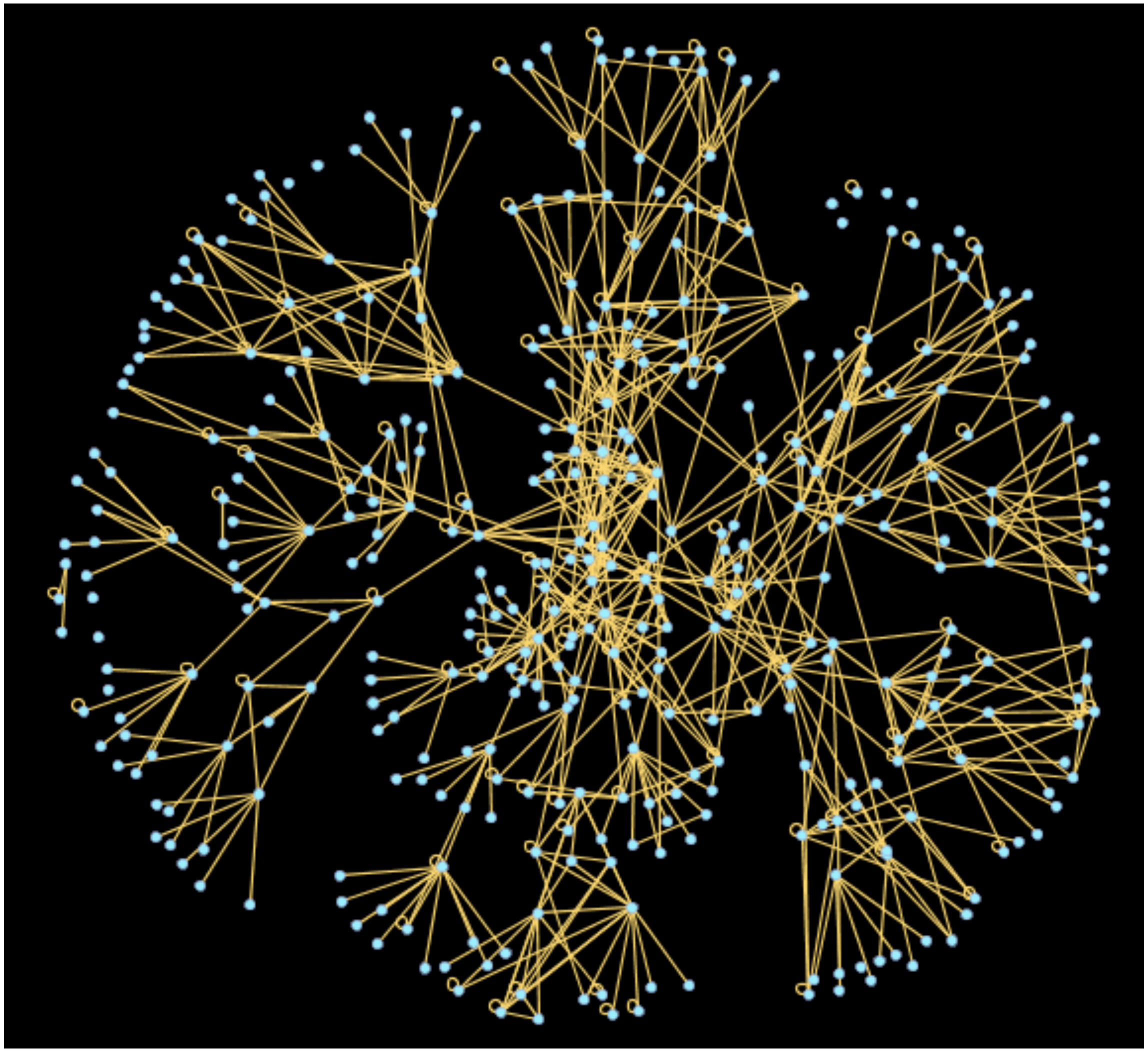
Edges are between individuals who sent
at least 6 email messages each way

450 users
median degree = 10, mean degree = 13
average shortest path = 3

Node properties specified:
degree
geographical location
position in organizational hierarchy

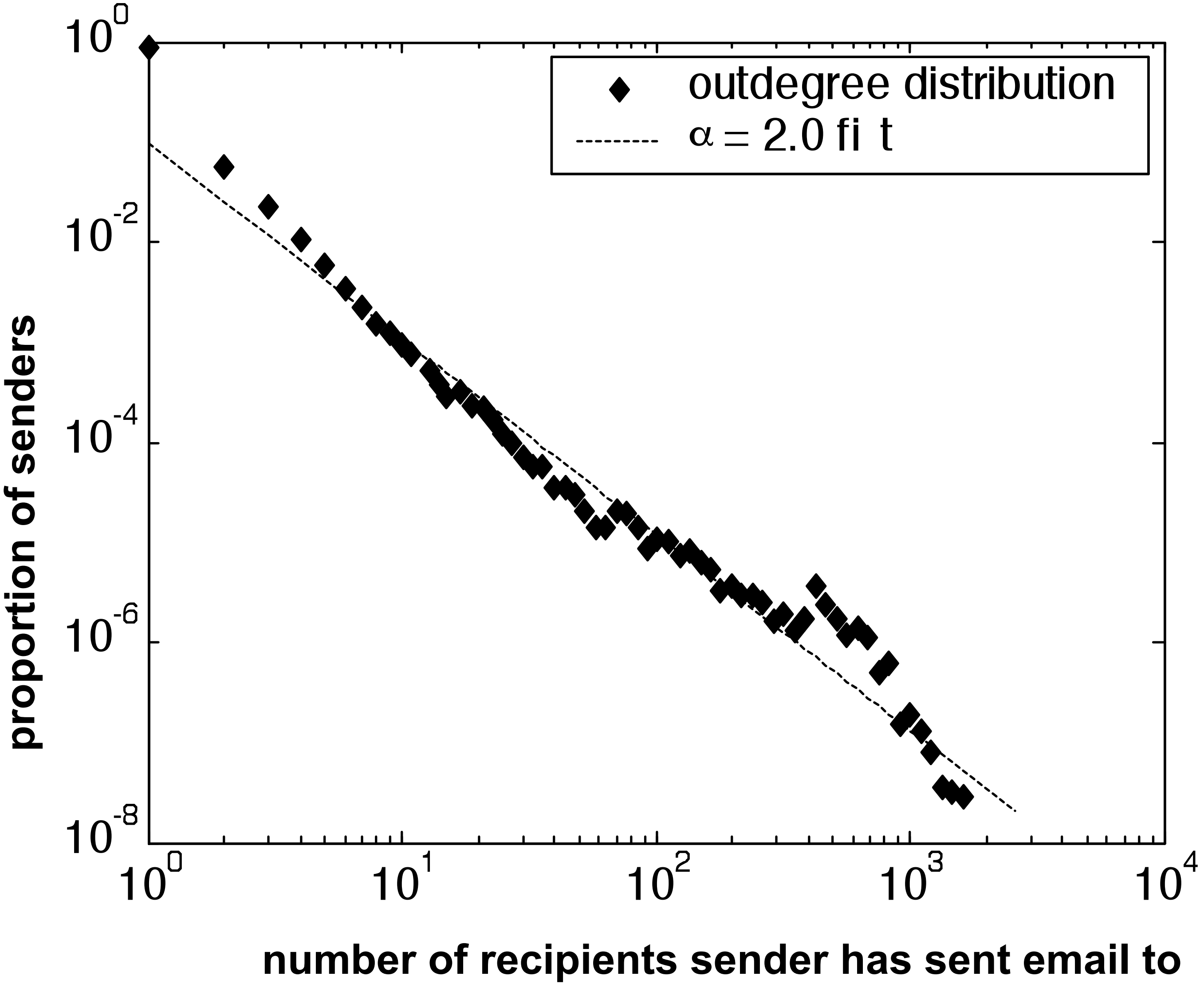
Can greedy strategies work?

the network otherwise known as sample.gdf



Strategy 1: High degree search

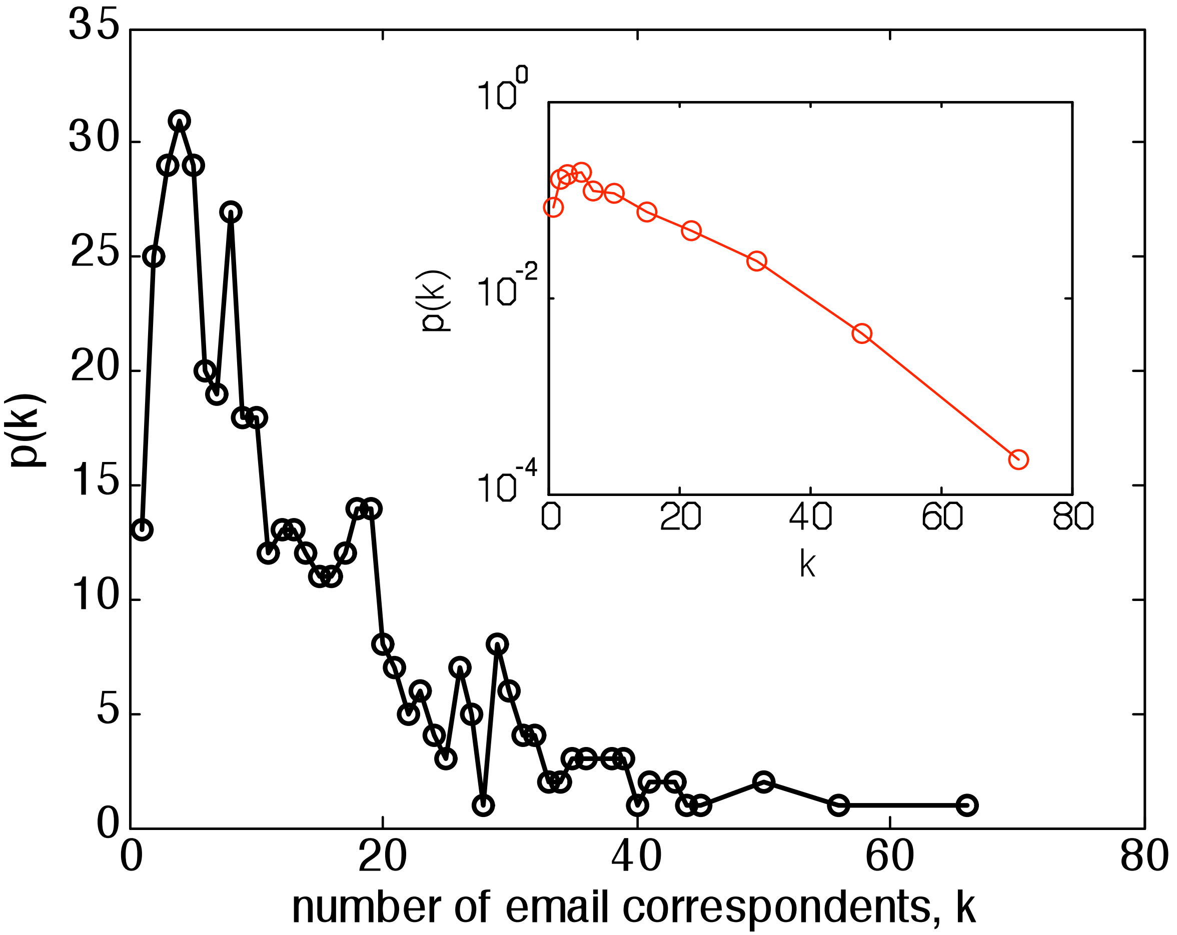
Power-law degree distribution of all senders of email passing through HP labs



Filtered network

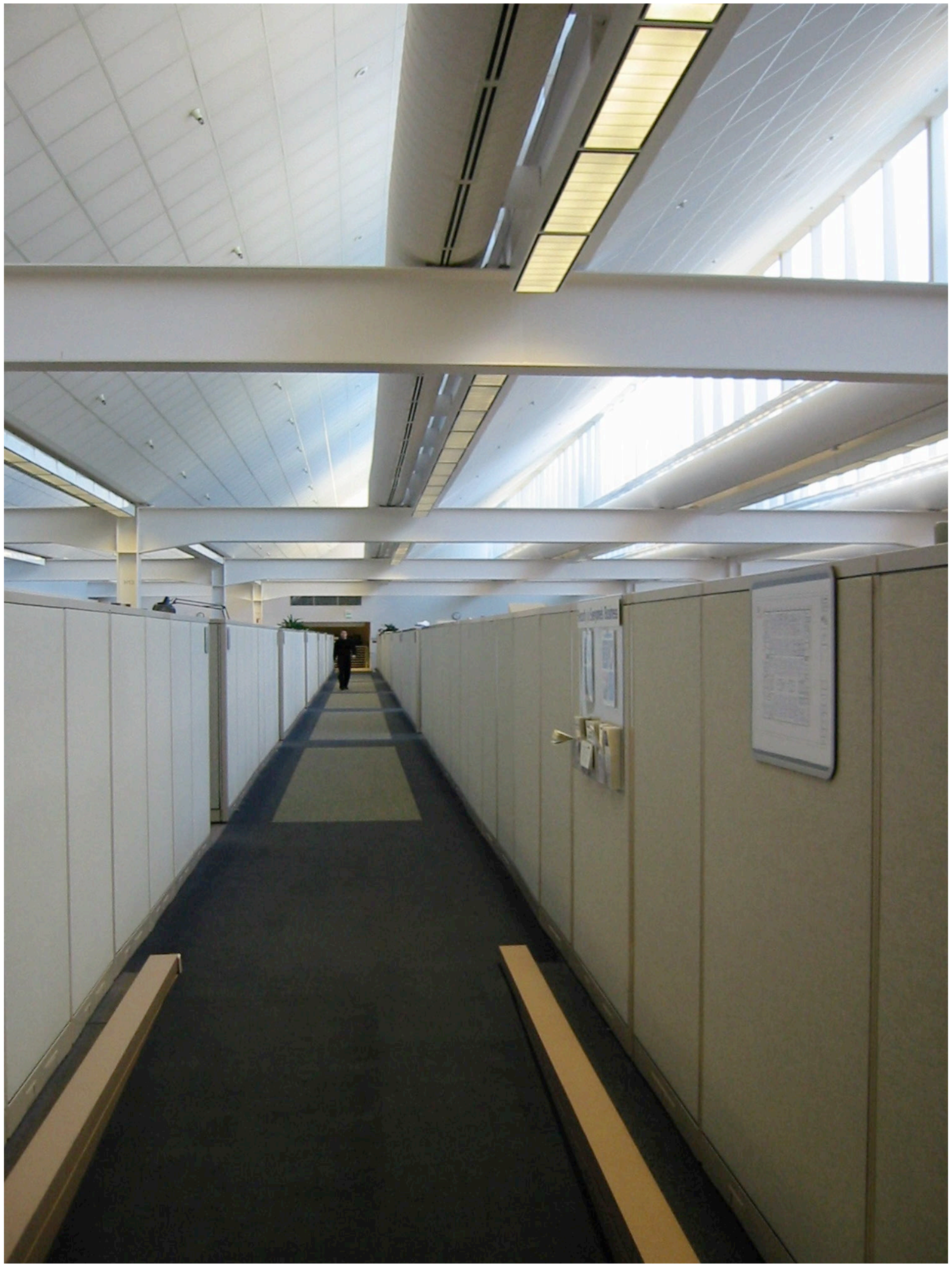
(at least 6 messages sent each way)

Degree distribution no longer power-law, but Poisson

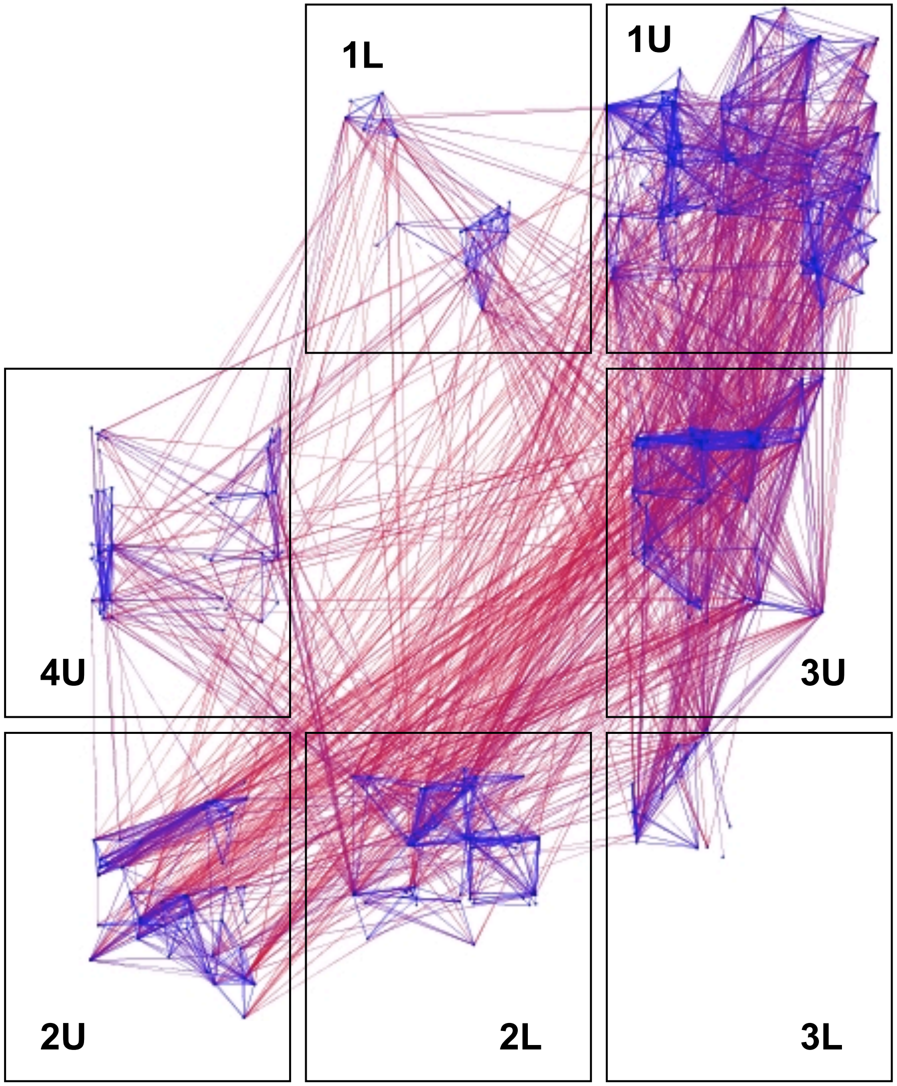


It would take 40 steps on average (median of 16) to reach a target!

Strategy 2:
Geography



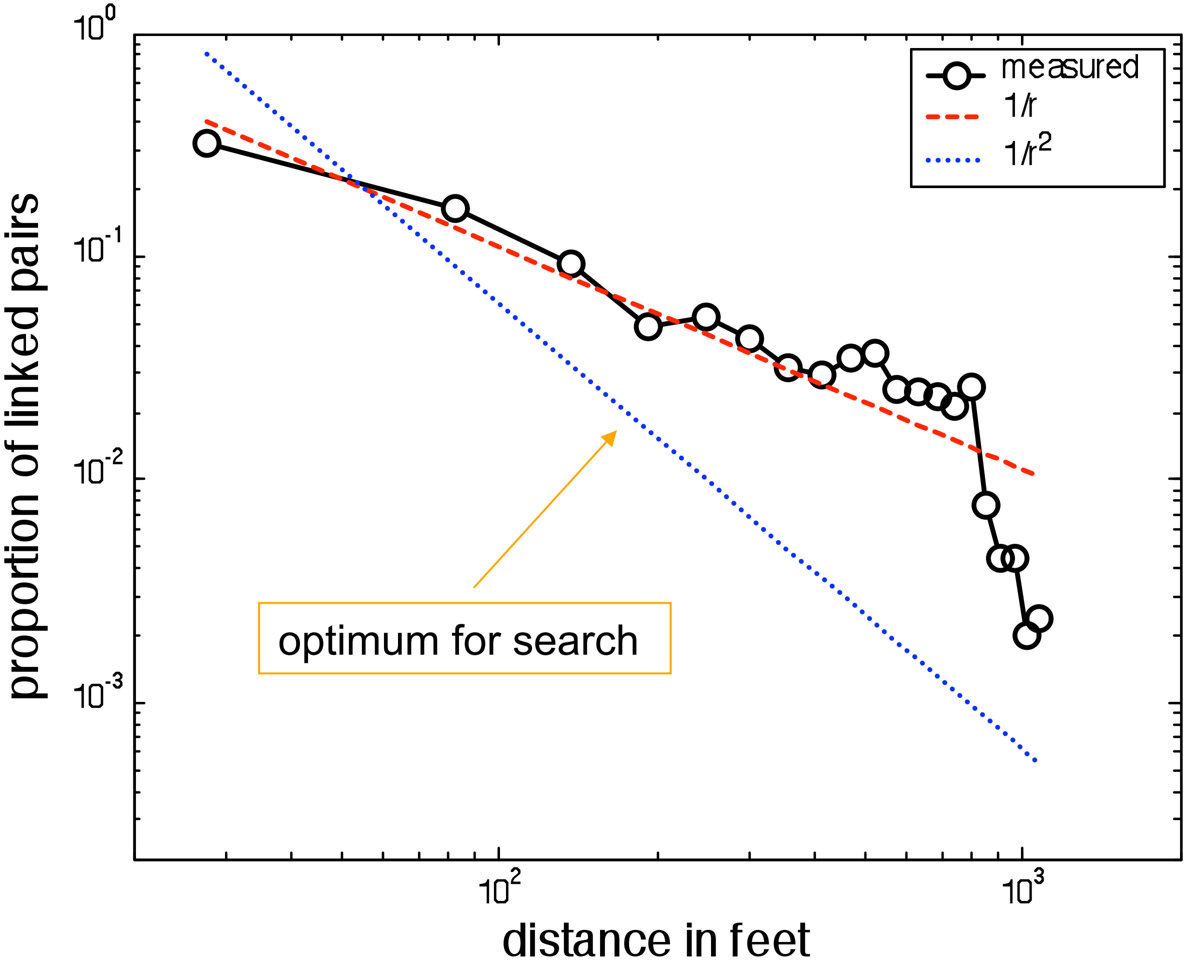
Communication across corporate geography



87 % of the
4000 links are
between individuals
on the same floor

source: Adamic and Adar, How to search a social network, Social Networks, 27(3), p.187-203, 2005.

Cubicle distance vs. probability of being linked



optimum for search

source: Adamic and Adar, How to search a social network, Social Networks, 27(3), p.187-203, 2005.

Livejournal

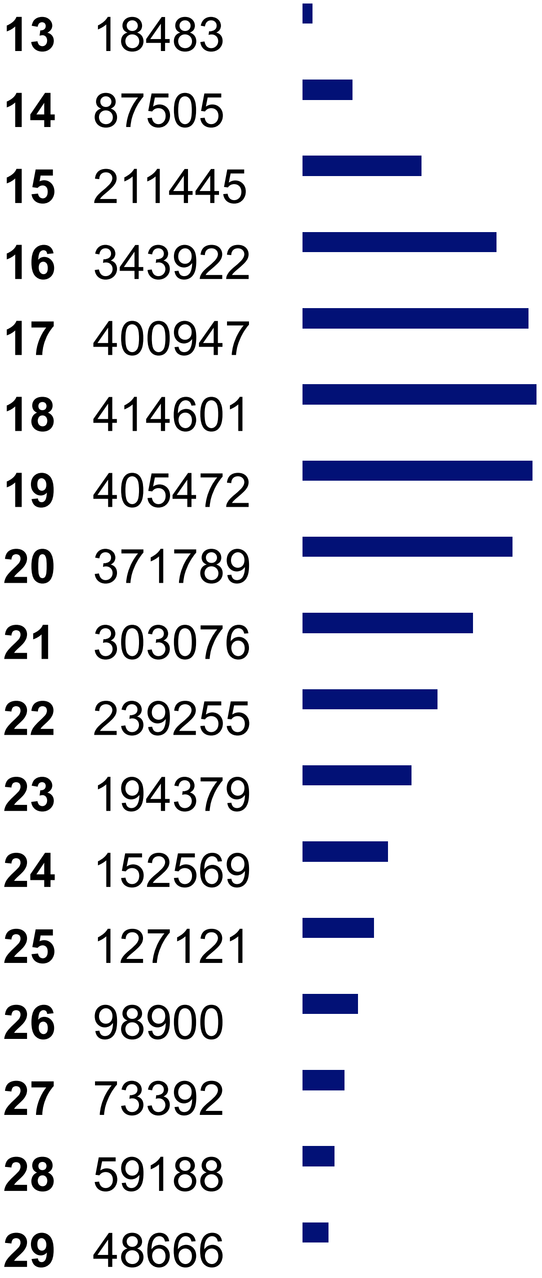
- LiveJournal provides an API to crawl the friendship network + profiles
  - friendly to researchers
  - great research opportunity

- basic statistics
  - **Users (stats from April 2006)**
    - How many users, and how many of those are active?
    - **Total accounts:** 9980558
    - **... active in some way:** 1979716
    - **... that have ever updated:** 6755023
    - **... updating in last 30 days:** 1300312
    - **... updating in last 7 days:** 751301
    - **... updating in past 24 hours:** 216581

Predominantly female & young demographic

- **Male:** 1370813 (32.4%)
- **Female:** 2856360 (67.6%)
- **Unspecified:** 1575389

Age distribution



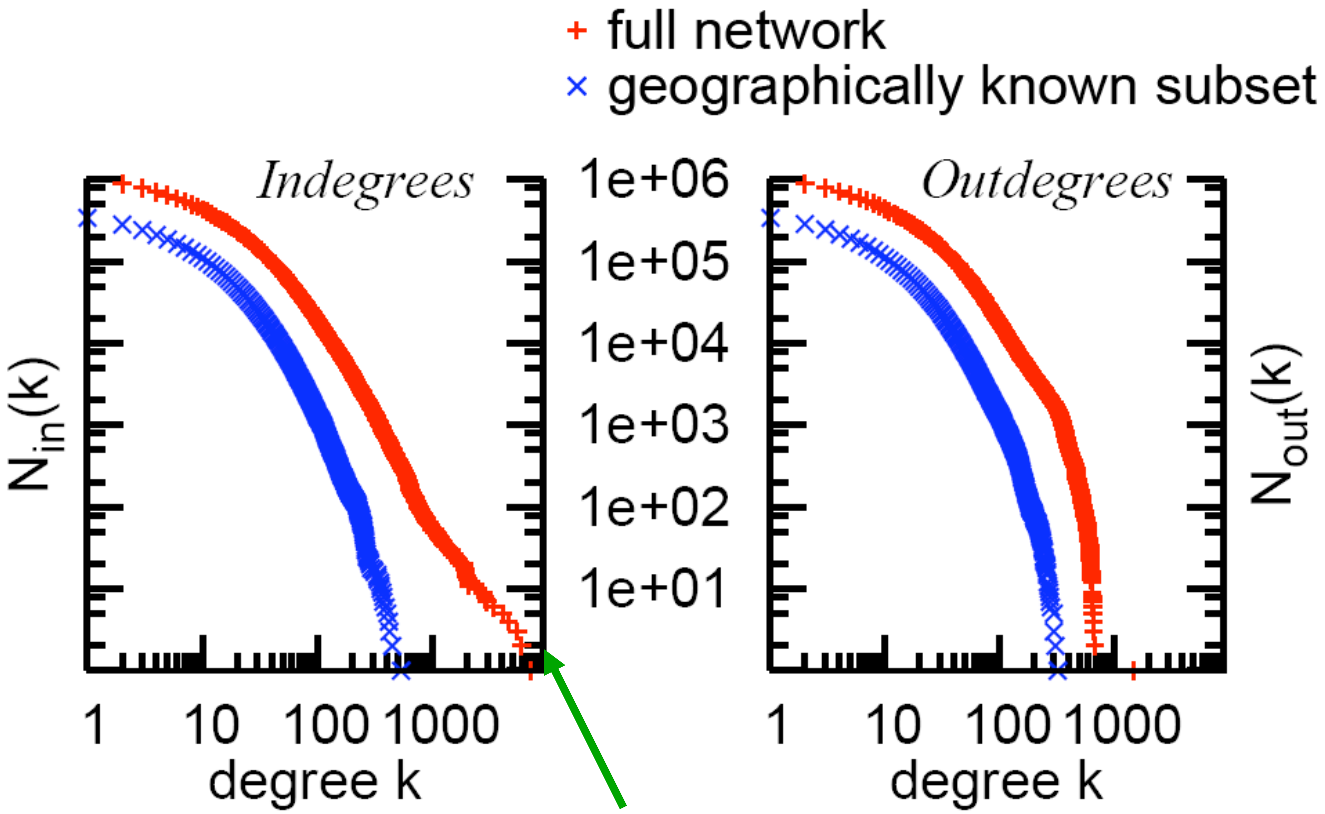
13	18483
14	87505
15	211445
16	343922
17	400947
18	414601
19	405472
20	371789
21	303076
22	239255
23	194379
24	152569
25	127121
26	98900
27	73392
28	59188
29	48666

Geographic Routing in Social Networks

- David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins (PNAS 2005)
- data used
  - Feb. 2004
  - 500,000 LiveJournal users with US locations
  - giant component (77.6%) of the network
  - clustering coefficient: 0.2

Degree distributions

- The broad degree distributions we’ve learned to know and love
  - but more probably lognormal than power law

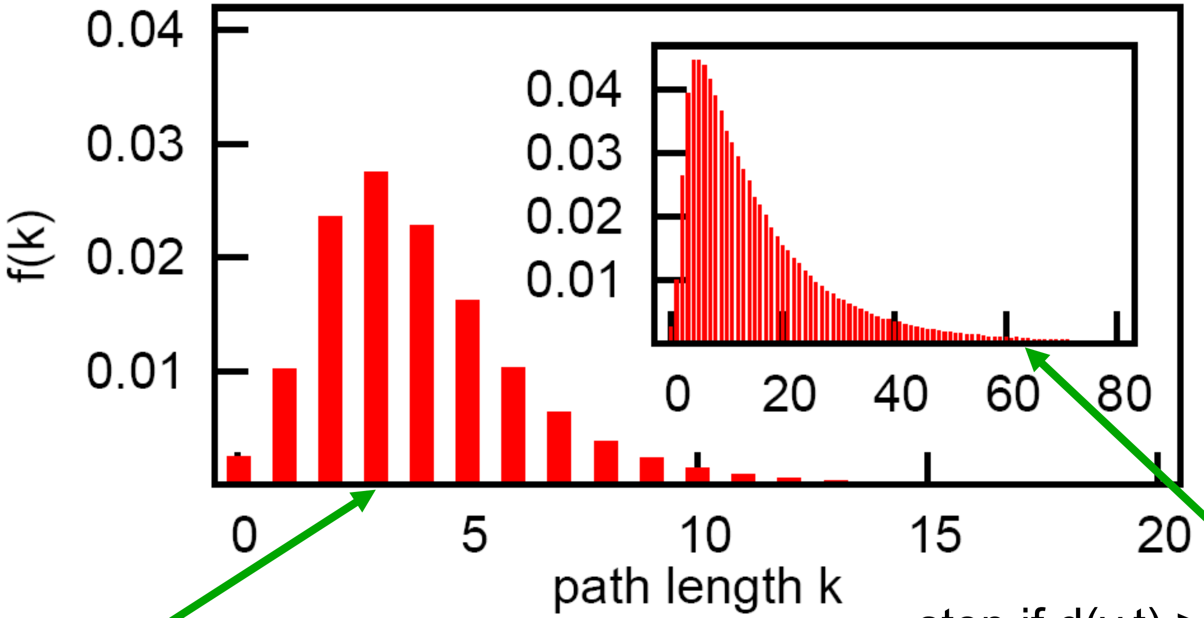


broader in degree than outdegree distribution

Source: http://www.tomkinshome.com/andrew/papers/science-blogs/pnas.pdf

Results of a simple greedy geographical algorithm

- Choose source *s* and target *t* randomly
- Try to reach target's city – not target itself
- At each step, the message is forwarded from the current message holder *u* to the friend *v* of u geographically closest to t



stop if d(v,t) > d(u,t)

13% of the chains are completed

stop if d(v,t) > d(u,t)

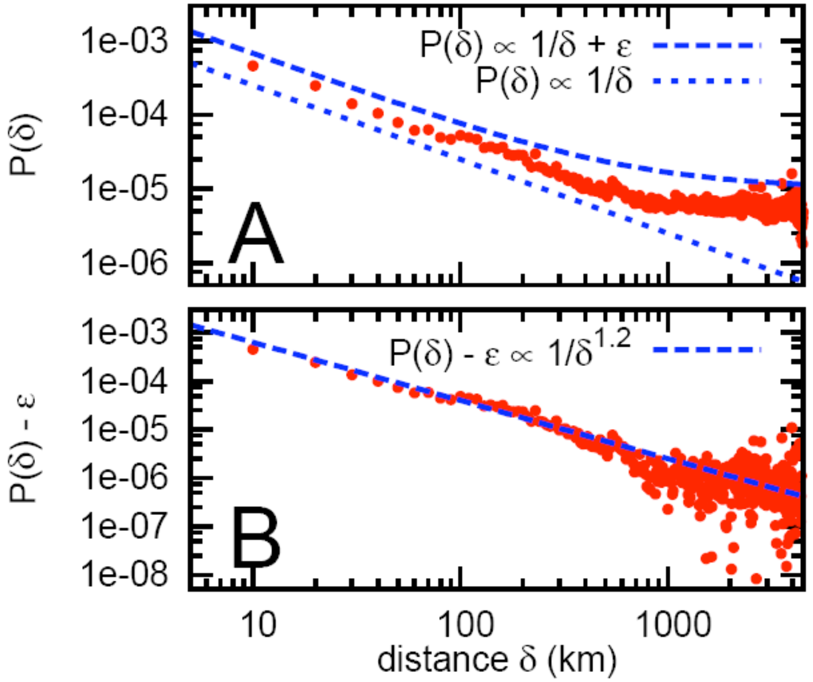
pick a neighbor at random in the same city if possible, else stop

80% of the chains are completed

Source: http://www.tomkinshome.com/andrew/papers/science-blogs/pnas.pdf

the geographic basis of friendship

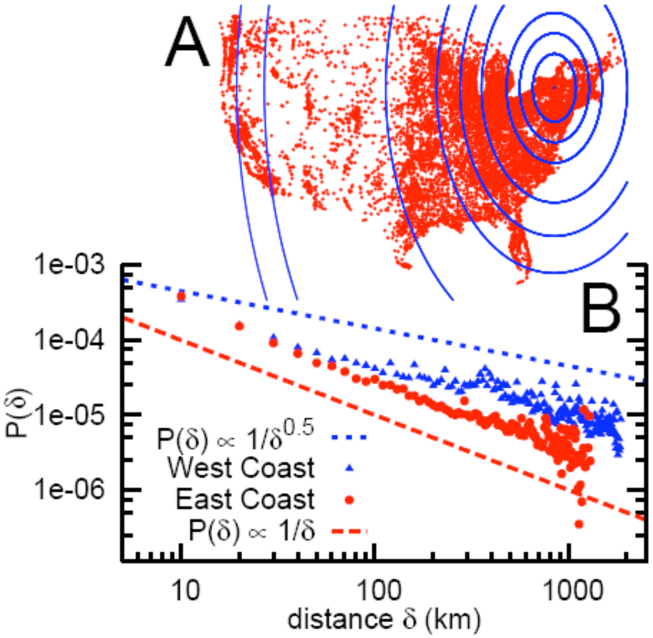
- δ = d(u,v) the distance between pairs of people
- The probability that two people are friends given their distance is equal to
  - $P(\delta) = \varepsilon + f(\delta)$, ε is a constant independent of geography
  - ε is 5.0 x 10^{-6} for LiveJournal users who are very far apart



Source: http://www.tomkinshome.com/andrew/papers/science-blogs/pnas.pdf

the geographic basis of friendship

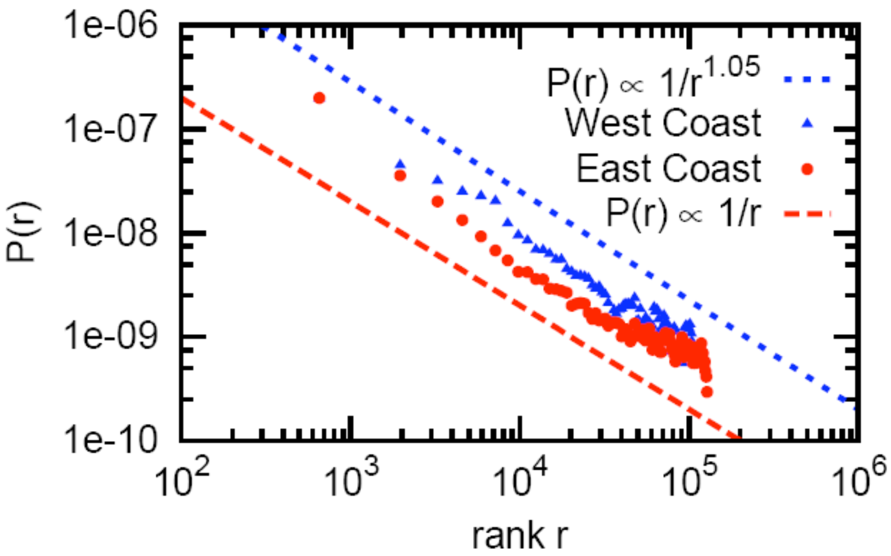
- The average user will have ~ 2.5 non-geographic friends
- The other friends (5.5 on average) are distributed according to an approximate 1/distance relationship
- But 1/d was proved not to be navigable by Kleinberg, so what gives?



Source: http://www.tomkinshome.com/andrew/papers/science-blogs/pnas.pdf

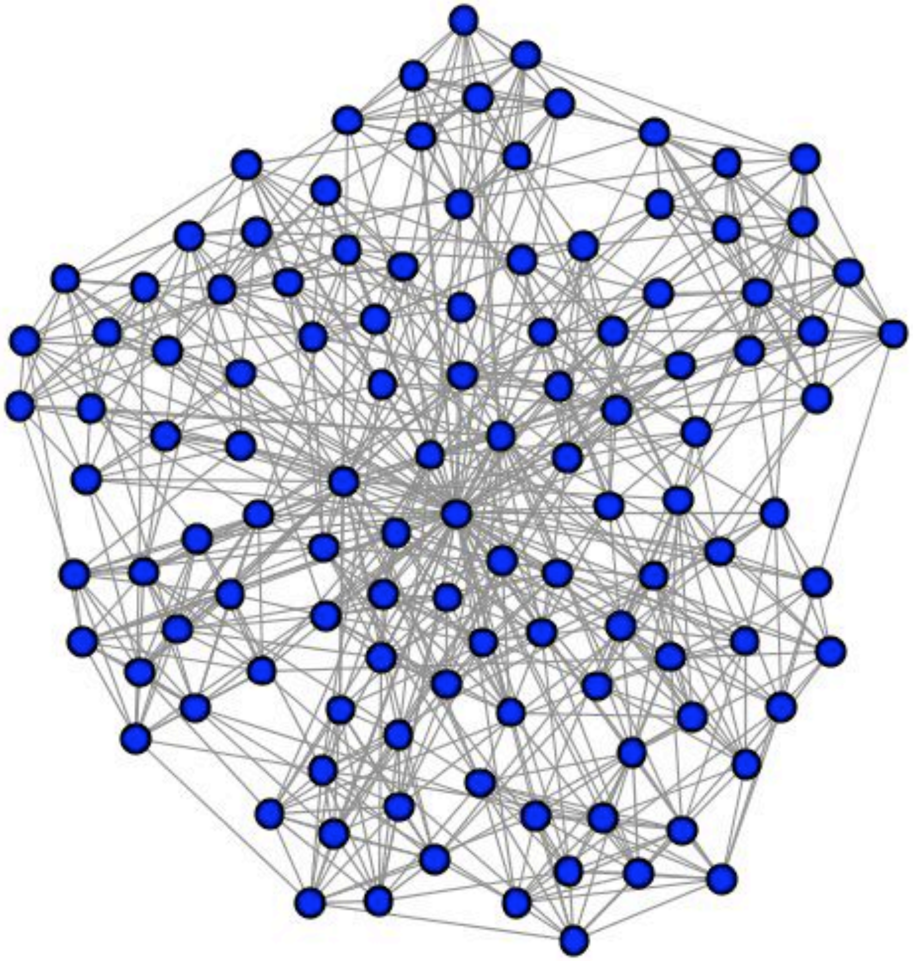
Navigability in networks of variable geographical density

- Kleinberg assumed a uniformly populated 2D lattice
- But population is far from uniform
- population networks and rank-based friendship
  - probability of knowing a person depends not on absolute distance but on relative distance (i.e. how many people live closer) Pr[u ->v] ~ 1/rank$_u$(v)

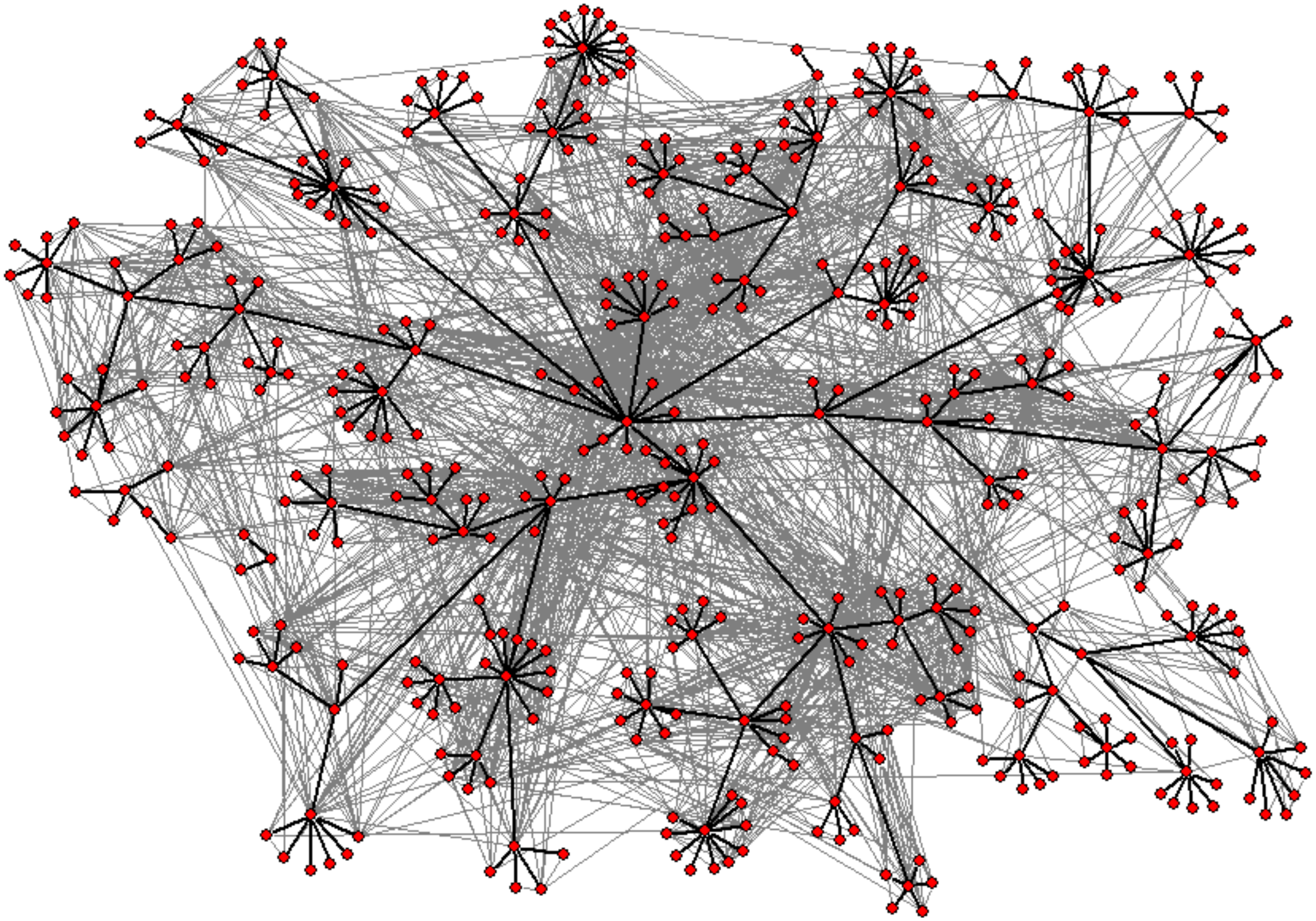


Source: http://www.tomkinshome.com/andrew/papers/science-blogs/pnas.pdf

what if we don’t have geography?

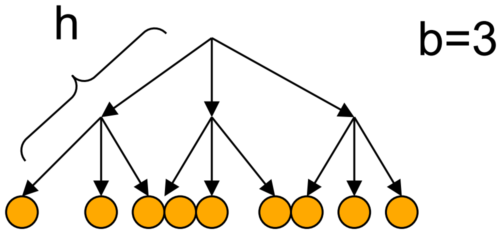


does community structure help?



review: hierarchical small world models

Individuals classified into a hierarchy,
h_{ij} = height of the least common ancestor.



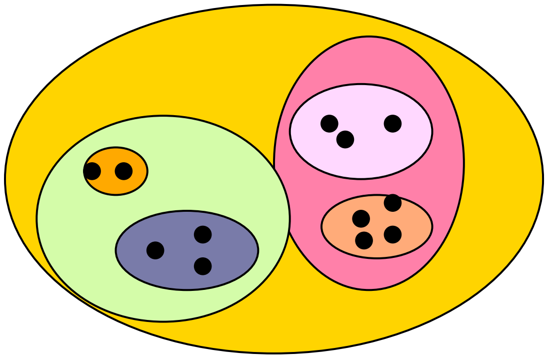
$$p_{ij} \sim b^{-\alpha h_{ij}}$$

e.g. state-county-city-neighborhood
industry-corporation-division-group

Theorem: If α = 1 and outdegree is polylogarithmic, can
s ~ O(log n)

Group structure models:
Individuals belong to nested groups
q = size of smallest group that v,w belong to

$$f(q) \sim q^{-\alpha}$$

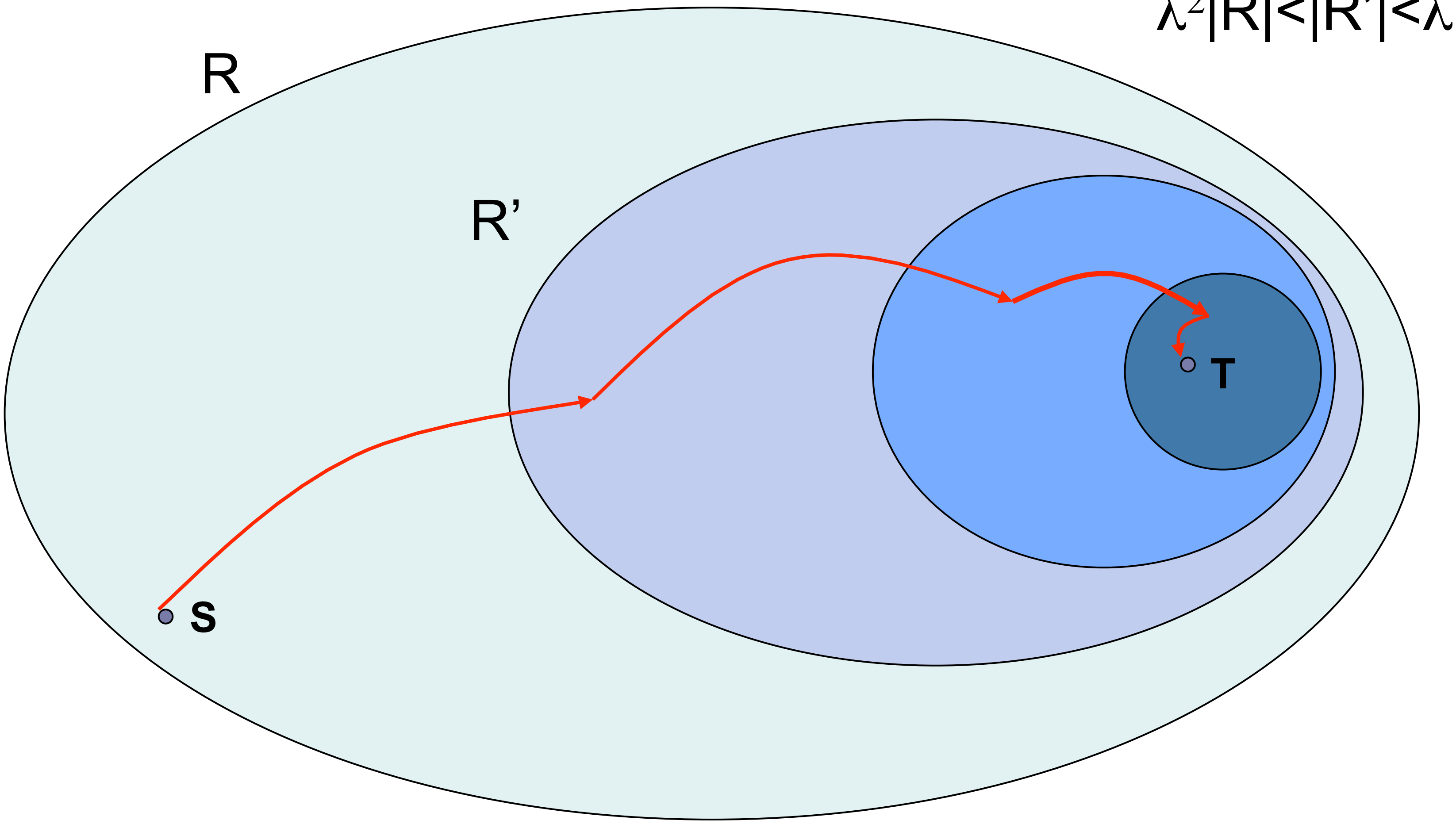


Theorem: If α = 1 and outdegree is polylogarithmic, can
s ~ O(log n)

Kleinberg, 'Small-World Phenomena and the Dynamics of Information', NIPS 14, 2001

Why search is fast in hierarchical topologies

$\lambda^2|R|<|R'|<\lambda|R|$

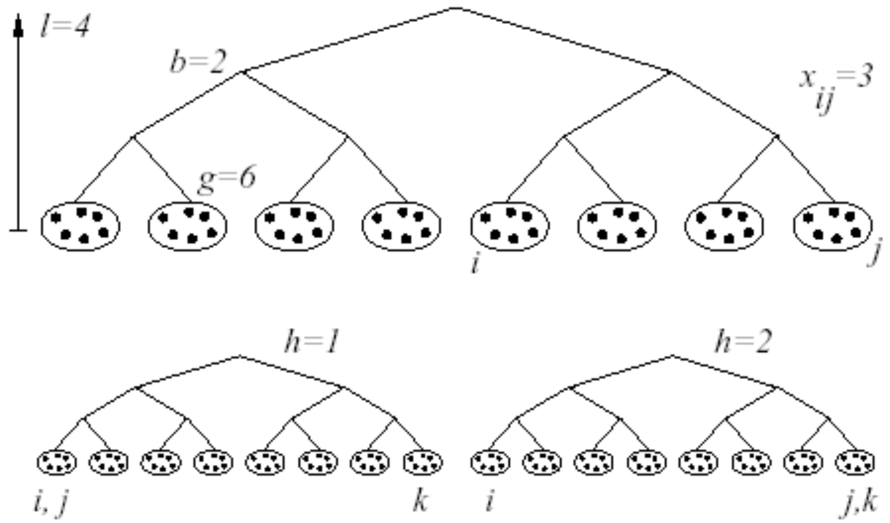


$k = c \log^2 n$

calculate probability that s fails to have a link in R’

hierarchical models with multiple hierarchies

individuals belong to hierarchically nested groups

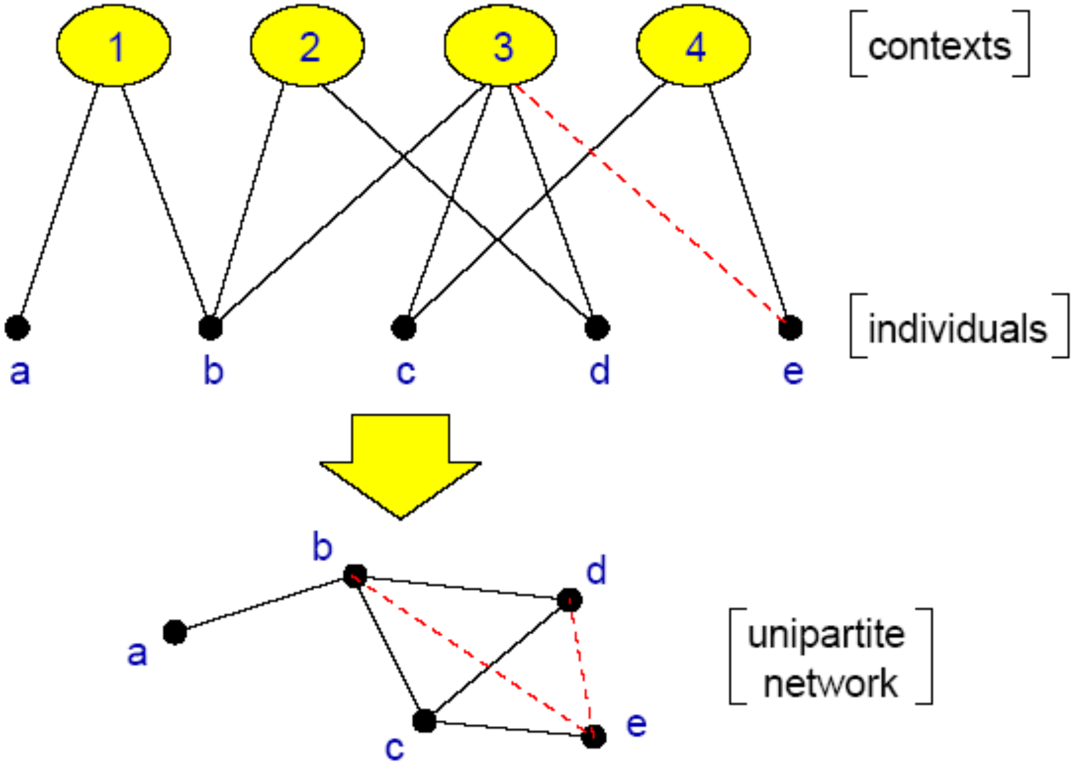


$p_{ij} \sim \exp(-\alpha\ x)$

multiple independent hierarchies h=1,2,..,H coexist corresponding to occupation, geography, hobbies, religion…

Source: Identity and Search in Social Networks: Duncan J. Watts, Peter Sheridan Dodds, and M. E. J. Newman; Science 17 May 2002 296: 1302-1305. < http://arxiv.org/abs/cond-mat/0205383v1 >

Social distance—Bipartite networks:



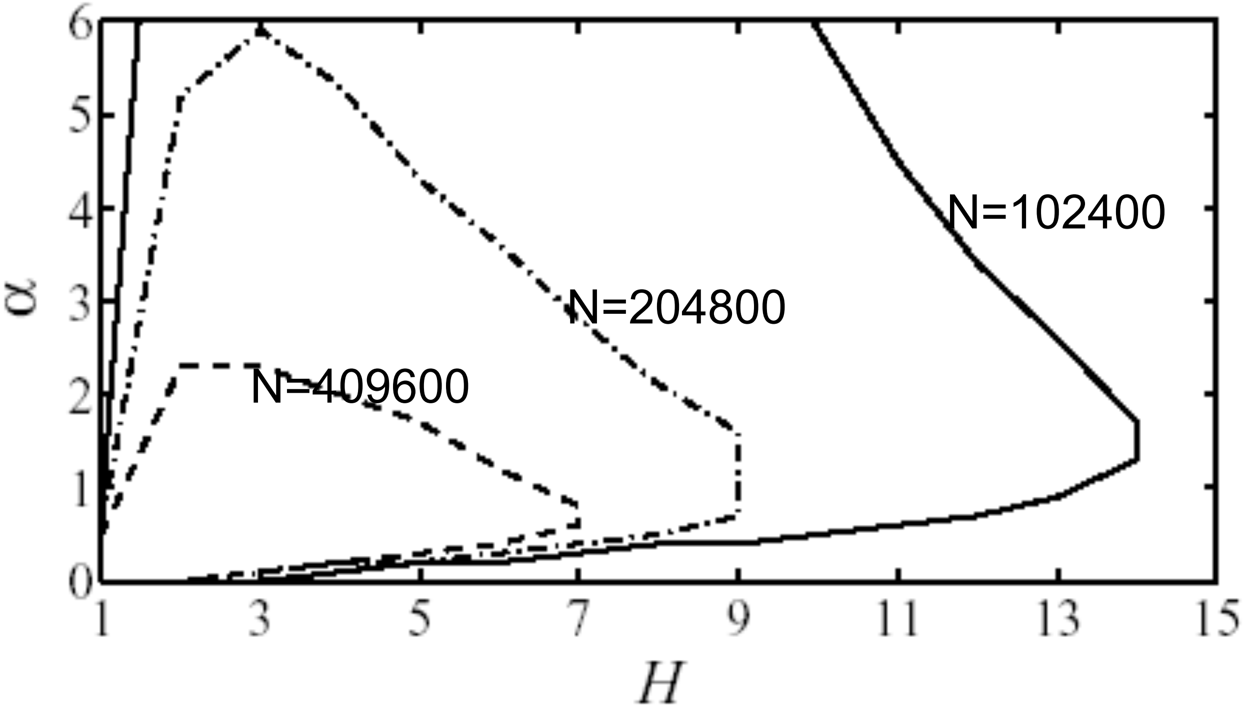
Source: Identity and Search in Social Networks: Duncan J. Watts, Peter Sheridan Dodds, and M. E. J. Newman; Science 17 May 2002 296: 1302-1305. < http://arxiv.org/abs/cond-mat/0205383v1 >

Identity and search in social networks

Watts, Dodds, Newman (2001)

Message chains fail at each node with probability p
Network is ‘searchable’ if a fraction r of messages reach the target

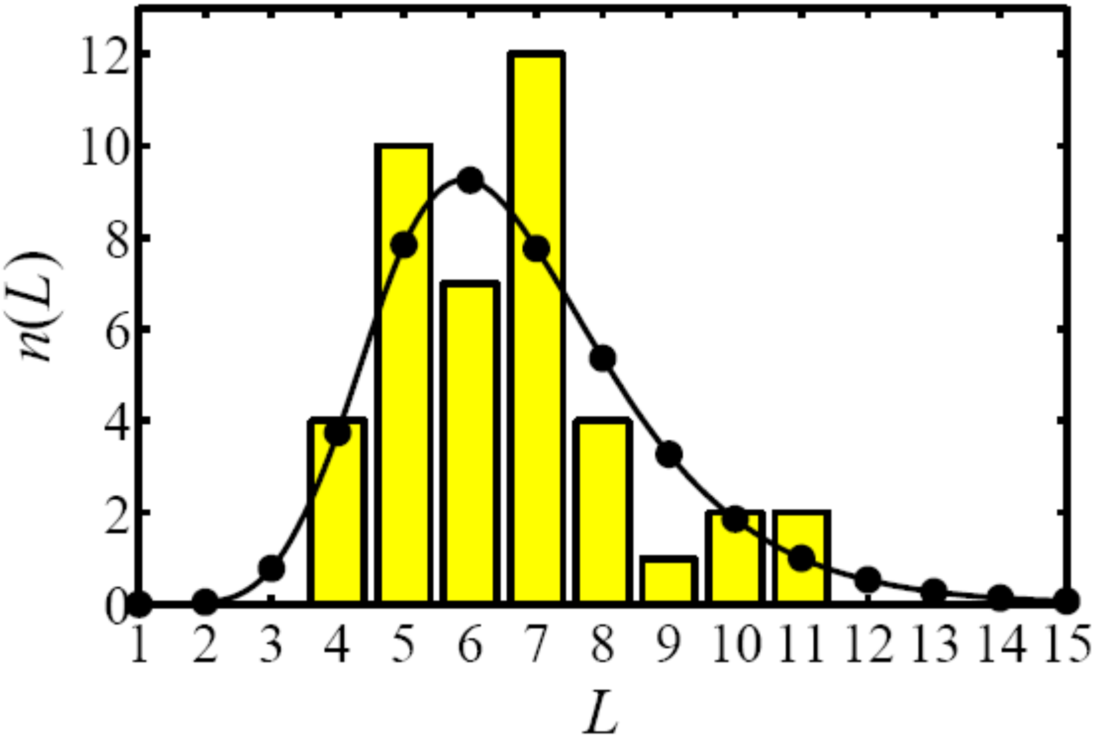
$$q = \left\langle (1-p)^L \right\rangle_L \geq r$$



Source: Identity and Search in Social Networks: Duncan J. Watts, Peter Sheridan Dodds, and M. E. J. Newman; Science 17 May 2002 296: 1302-1305. < http://arxiv.org/abs/cond-mat/0205383v1 >

Small World Model, Watts et al.

Fits Milgram's data well



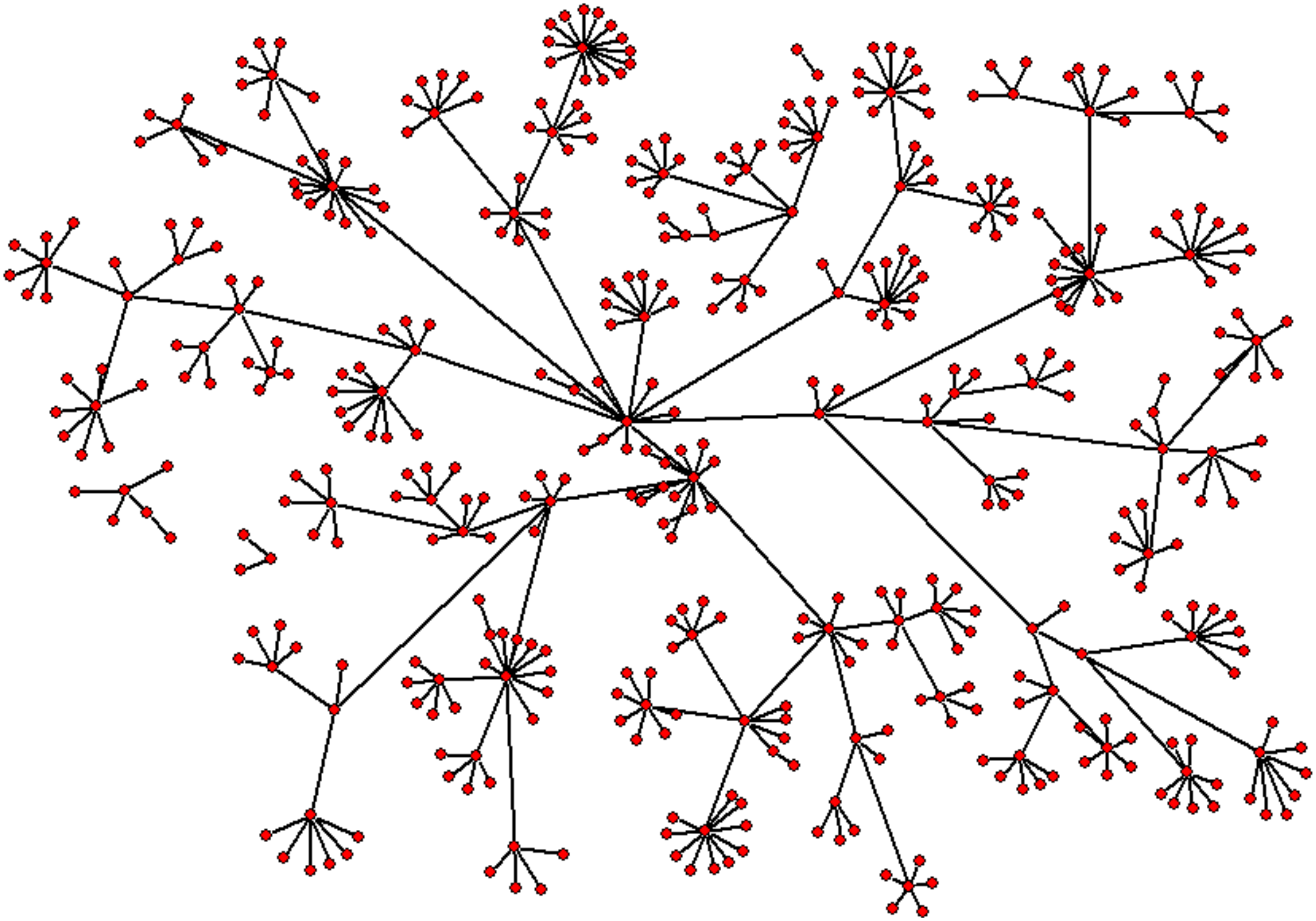
Model parameters:
N = 10^8
z = 300
g = 100
b = 10
α= 1, H = 2

L_{model}= 6.7
L_{data} = 6.5

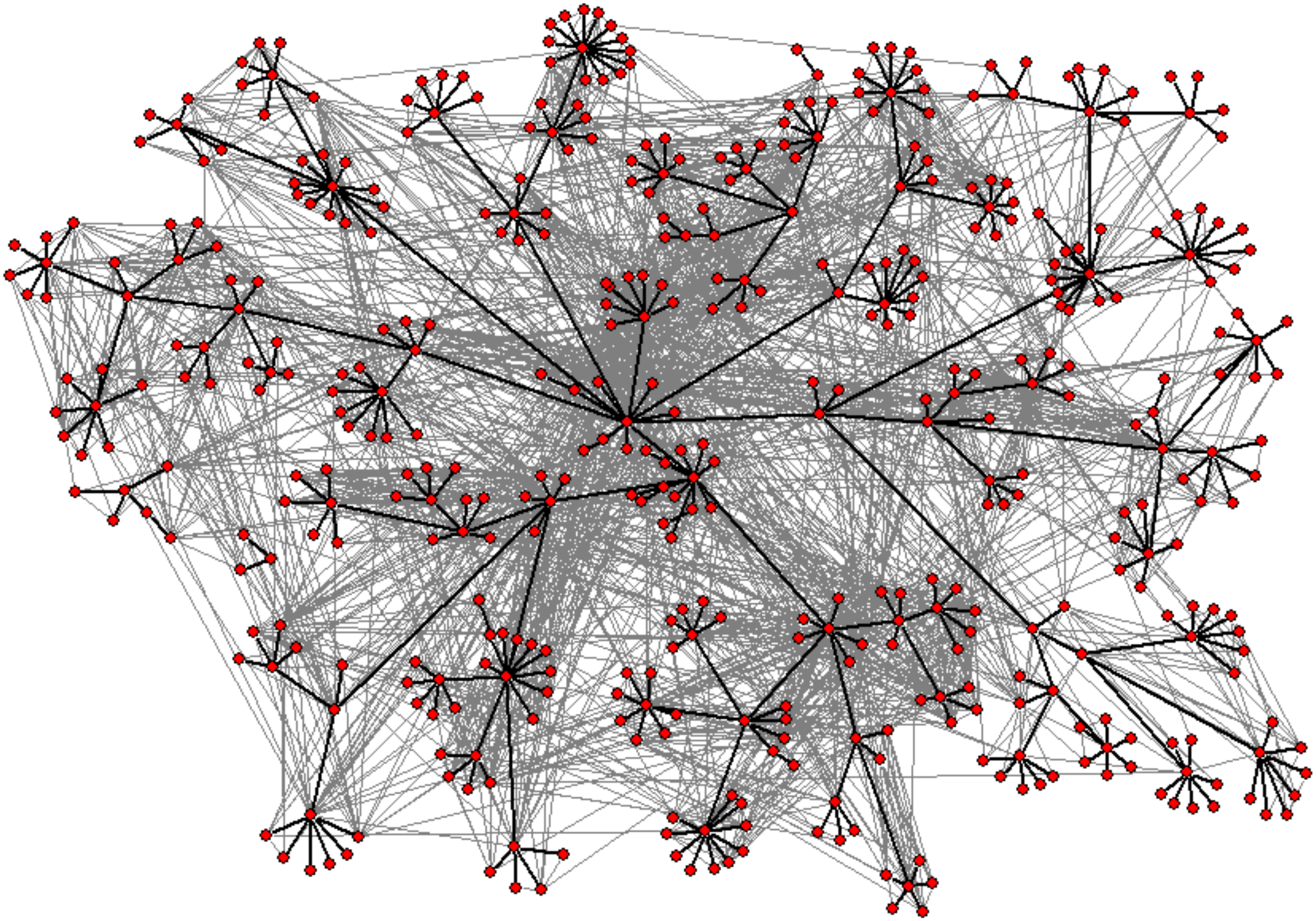
more slides on this:

http://www.aladdin.cs.cmu.edu/workshops/wsa/papers/dodds-2004-04-10search.pdf

does it work in practice? back to HP Labs: Organizational hierarchy

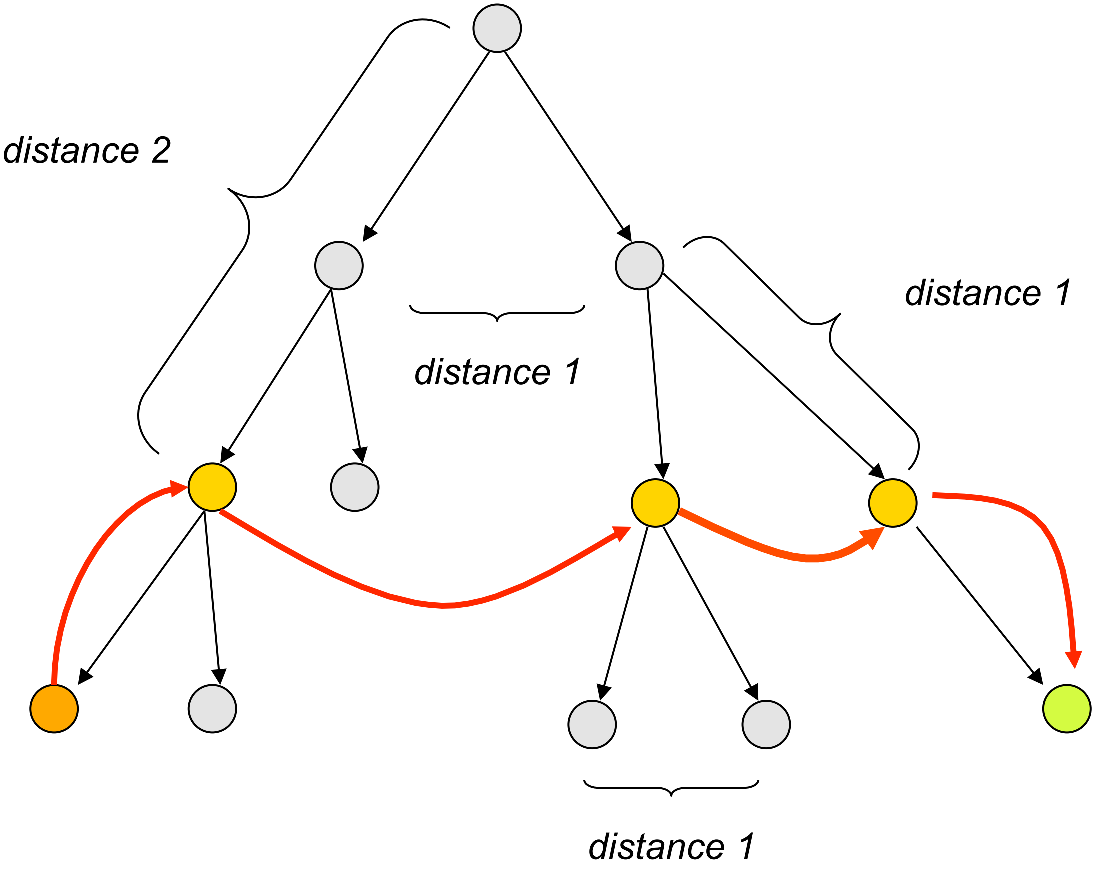


Email correspondence superimposed on the organizational hierarchy



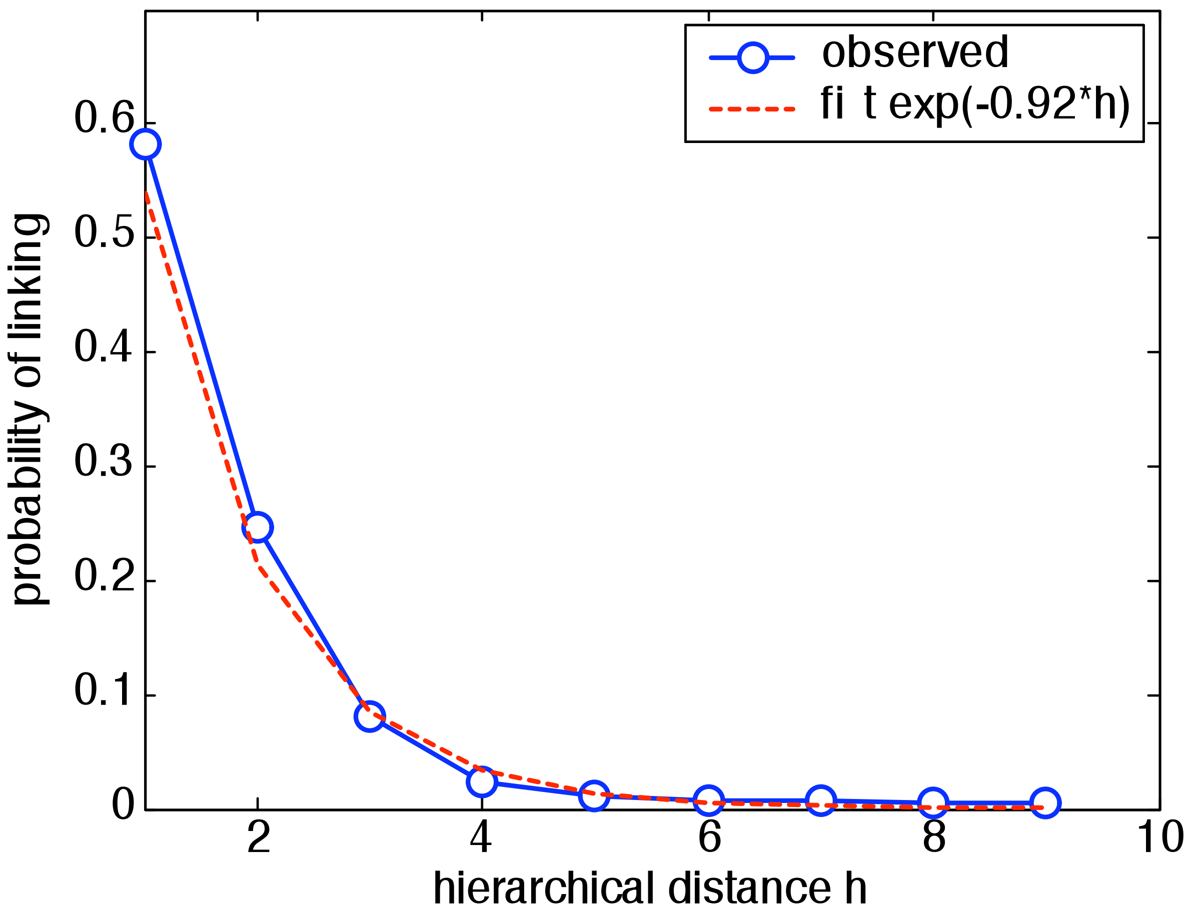
source: Adamic and Adar, How to search a social network, Social Networks, 27(3), p.187-203, 2005.

Example of search path



hierarchical distance = 5
search path distance = 4

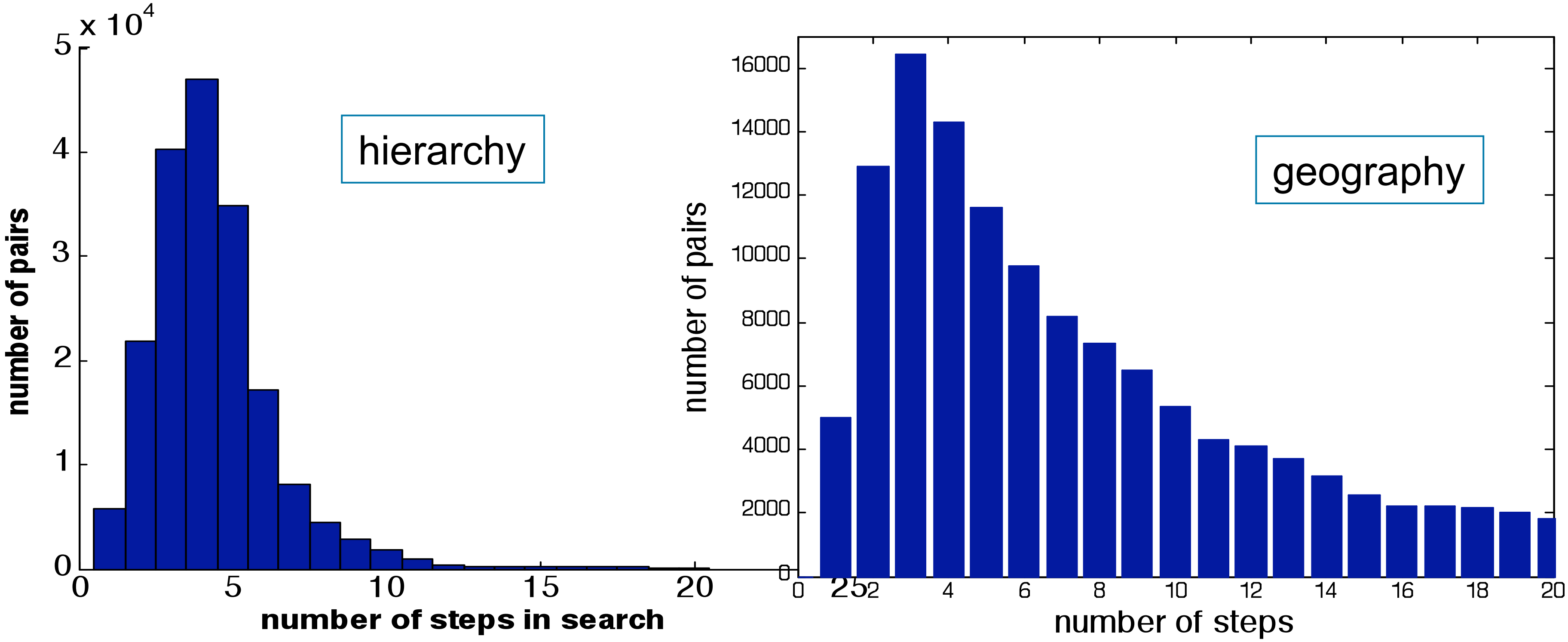
Probability of linking vs. distance in hierarchy



in the 'searchable' regime: $0 < \alpha < 2$ (Watts, Dodds, Newman 2001)

Results

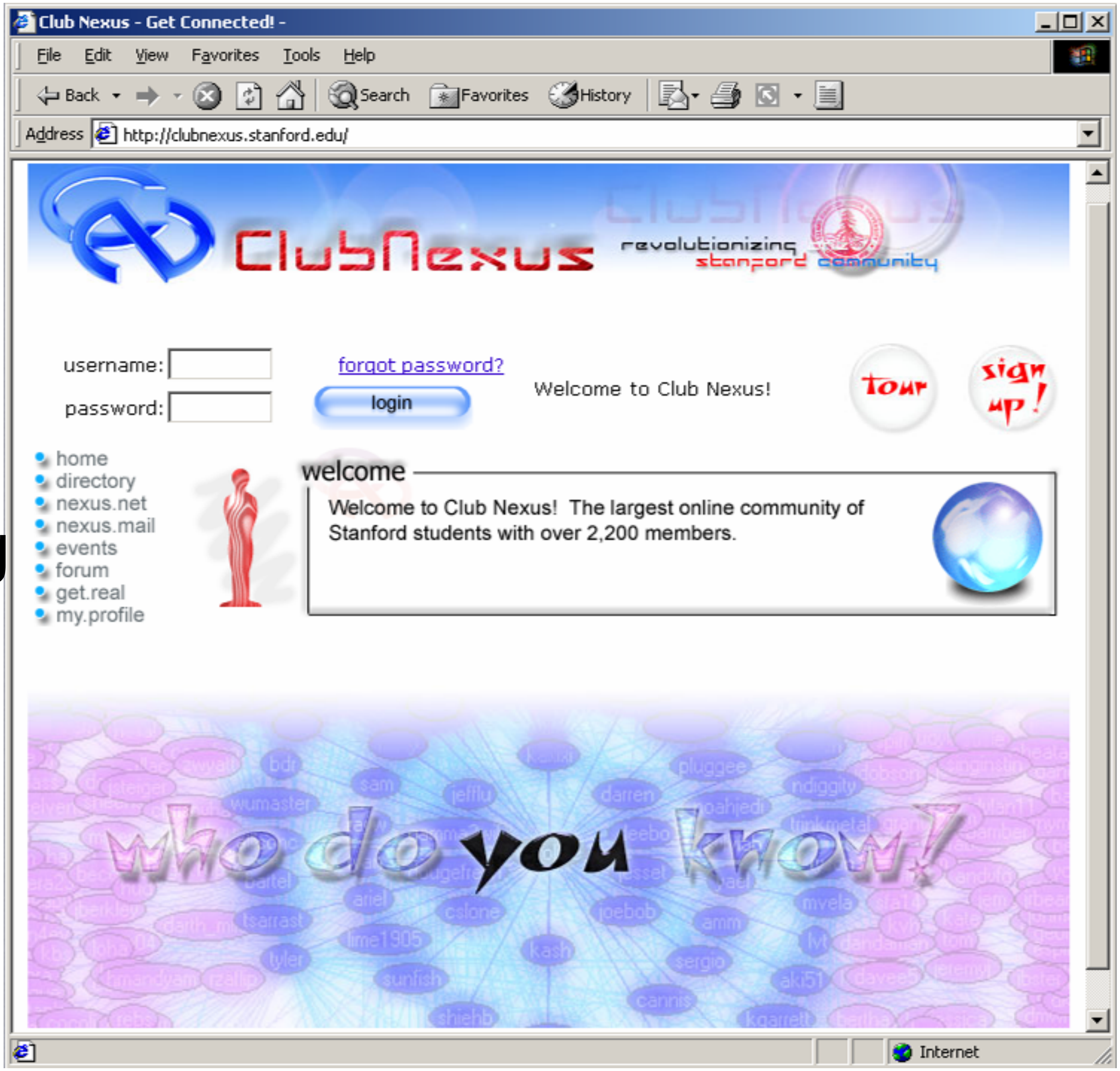
distance	hierarchy	geography	geodesic	org	random
median	4	7	3	6	28
mean	5.7 (4.7)	12	3.1	6.1	57.4



source: Adamic and Adar, How to search a social network, Social Networks, 27(3), p.187-203, 2005.

Expt 2

Searching a social networking website



Club Nexus - Get Connected! -

Address http://clubnexus.stanford.edu/

username:

password:

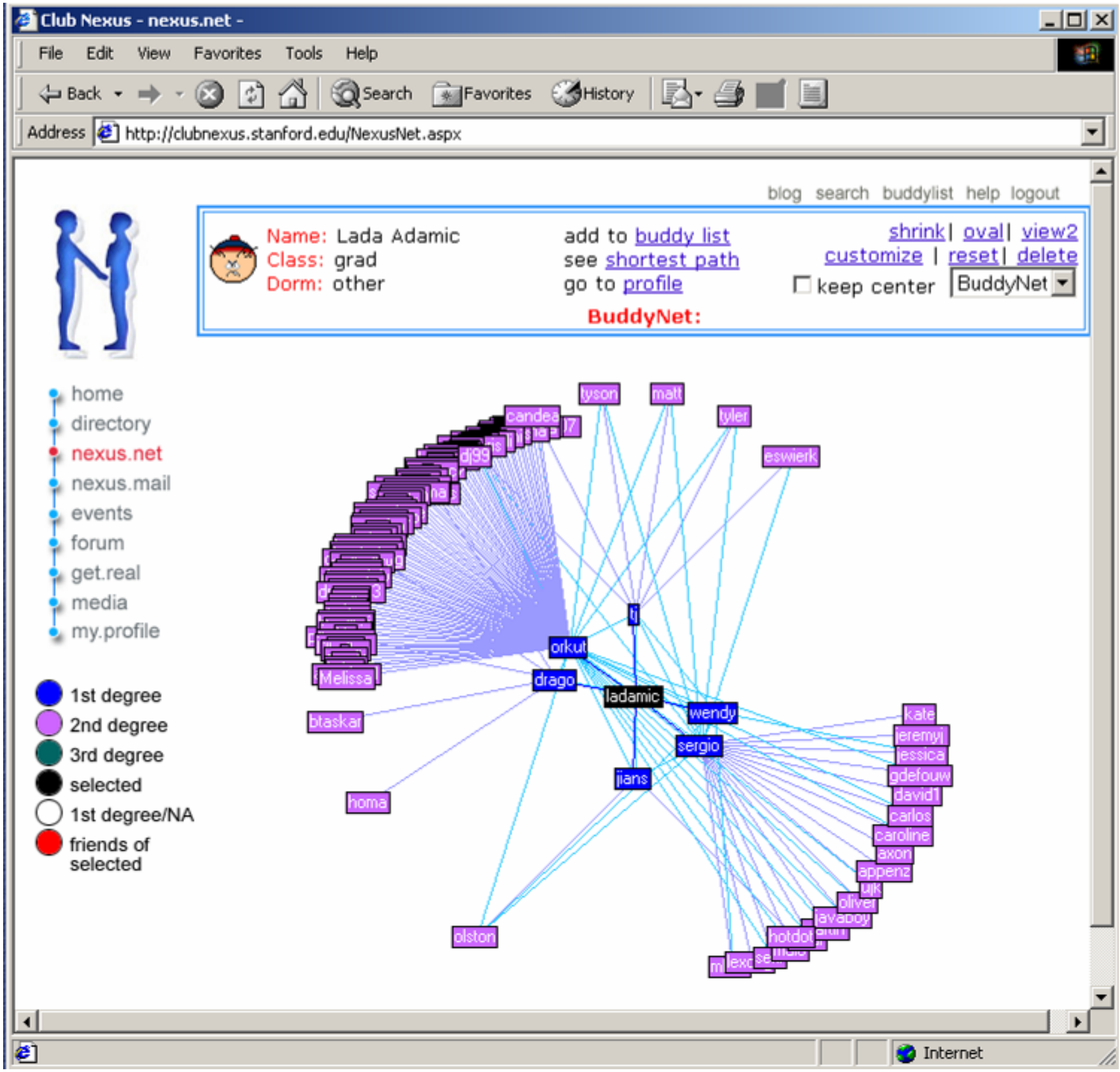
forgot password?

Welcome to Club Nexus!

welcome

Welcome to Club Nexus! The largest online community of Stanford students with over 2,200 members.

Source: ClubNexus - Orkut Buyukkokten, Tyler Ziemann



Address http://clubnexus.stanford.edu/NexusNet.aspx

Name: Lada Adamic
Class: grad
Dorm: other

□ keep center BuddyNet

BuddyNet:

Source: ClubNexus - Orkut Buyukkokten, Tyler Ziemann

Profiles:

status (UG or G)
year
major or **department**
residence
gender

Personality	(choose 3 exactly):
you	funny, kind, weird, …
friendship	honesty/trust, common interests, commitment, …
romance	- “ -
freetime	socializing, getting outside, reading, …
support	unconditional accepters, comic-relief givers, eternal optimists

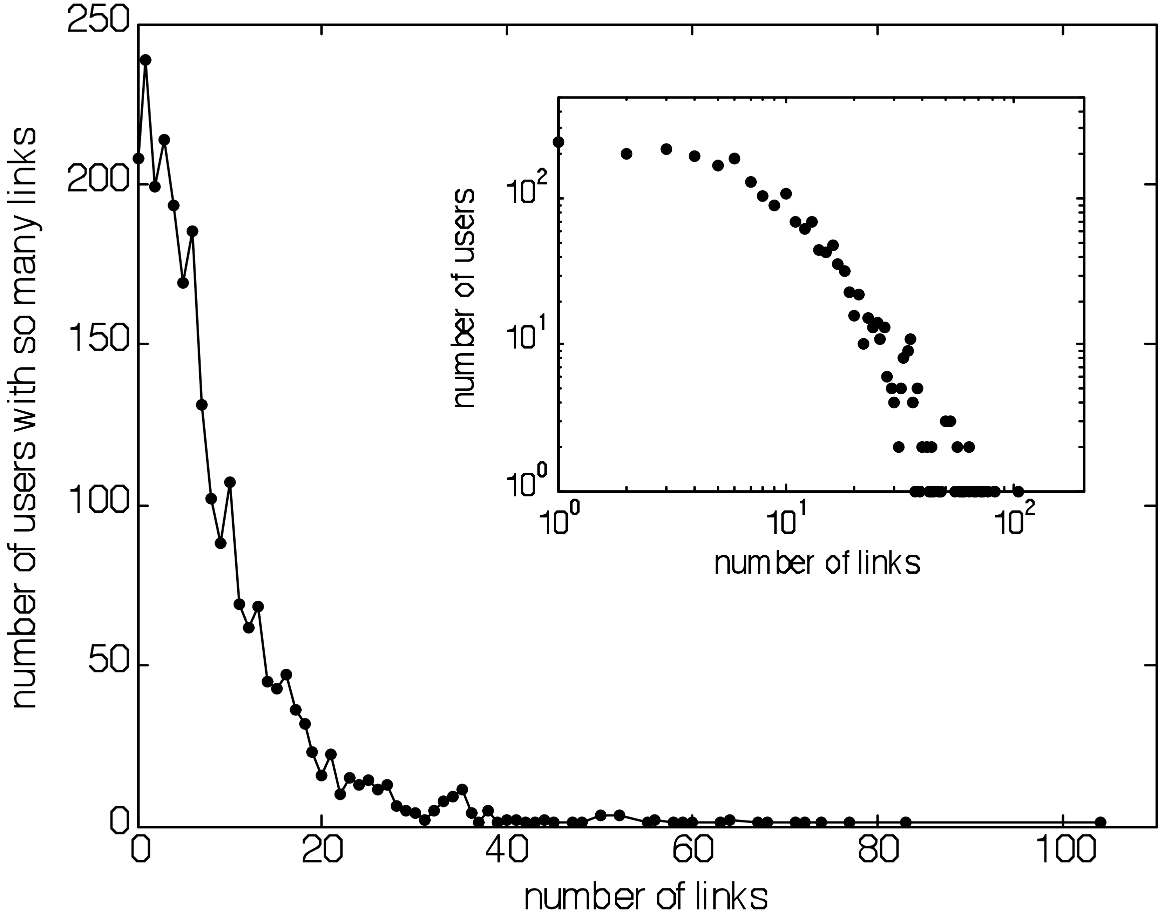
Interests	(choose as many as apply)
books	mystery & thriller, science fiction, romance, …
movies	western, biography, horror, …
music	folk, jazz, techno, …
social activities	ballroom dancing, barbecuing, bar-hopping, …
land sports	soccer, tennis, golf, …
water sports	sailing, kayaking, swimming, …
other sports	ski diving, weightlifting, billiards, …

Differences between data sets

HP labs email network	**Online community**
• complete image of communication network	• partial information of social network
• affinity not reflected	• only friends listed

Degree Distribution for Nexus Net

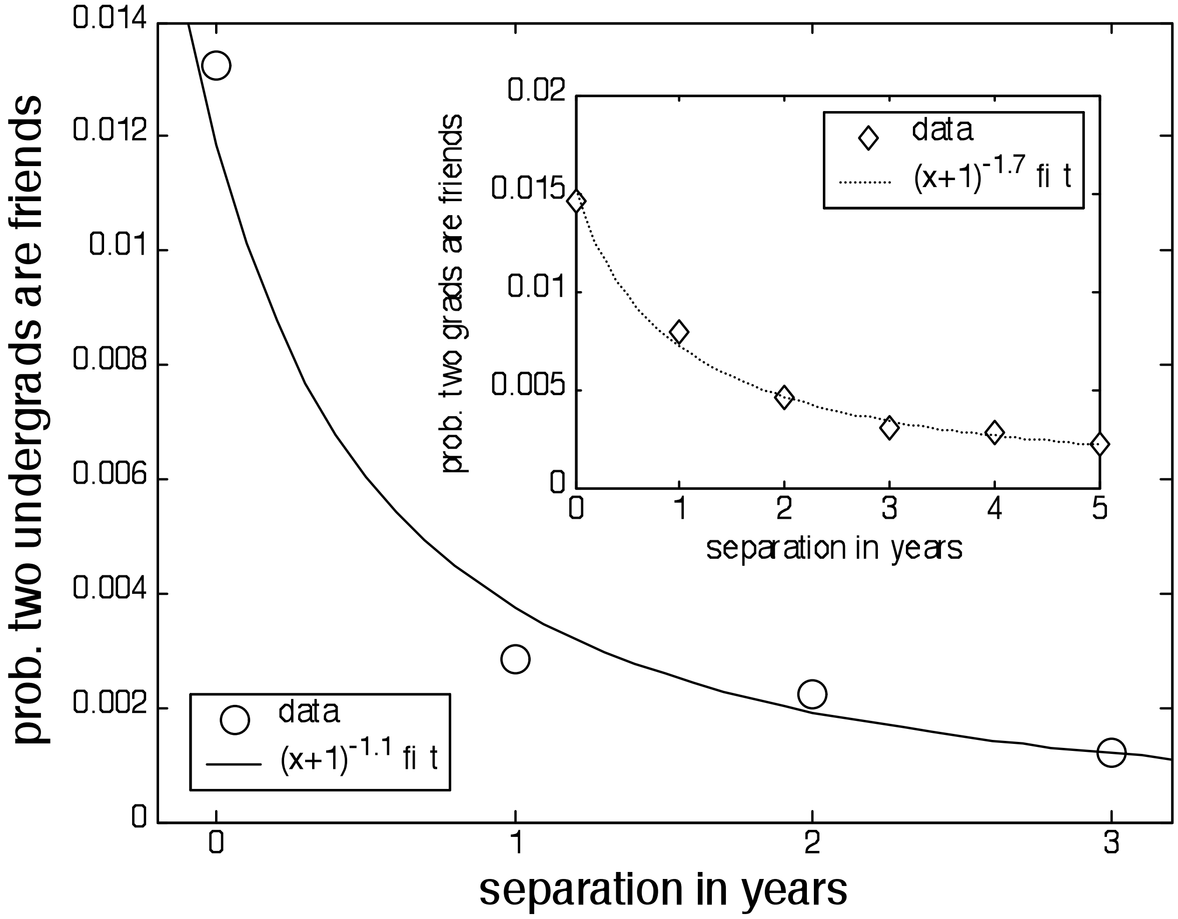
2469 users, average degree 8.2



source: Adamic and Adar, How to search a social network, Social Networks, 27(3), p.187-203, 2005.

Problem: how to construct hierarchies?

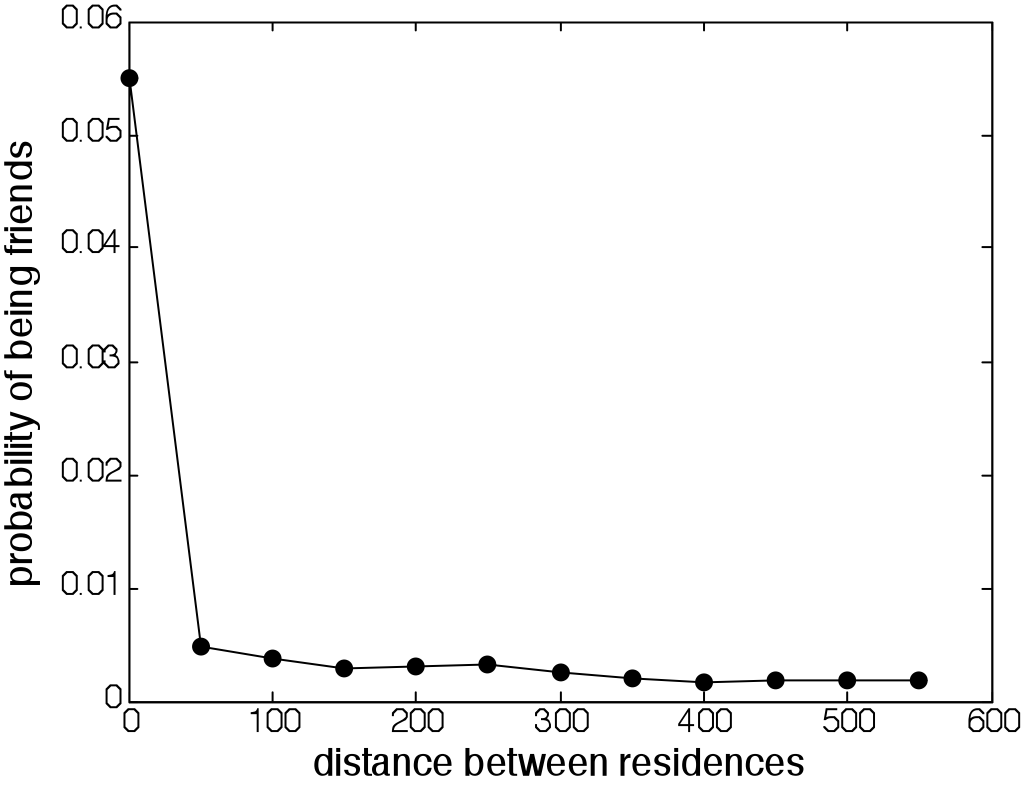
Probability of linking by separation in years



source: Adamic and Adar, **How to search a social network**, Social Networks, 27(3), p.187-203, 2005.

Hierarchies not useful for other attributes:

Geography



Other attributes: major, sports, freetime activities, movie preferences…

source: Adamic and Adar, **How to search a social network**, Social Networks, 27(3), p.187-203, 2005.

Strategy using user profiles

prob. two undergrads are friends (consider simultaneously)

- both undergraduate, both graduate, or one of each
- same or different year
- both male, both female, or one of each
- same or different residences
- same or different major/department

Results

strategy	median	mean
random	133	390
high degree	39	137
profile	21	53

With an attrition rate of 25%, 5% of the messages get through at an average of 4.8 steps,
=> hence network is *barely* searchable

conclusions

- Individuals associate on different levels into groups.
- Group structure facilitates decentralized search using social ties.
- Hierarchy search faster than geographical search
- A fraction of ‘important’ individuals are easily findable
- Humans may be more resourceful in executing search tasks:
  - making use of weak ties
  - using more sophisticated strategies