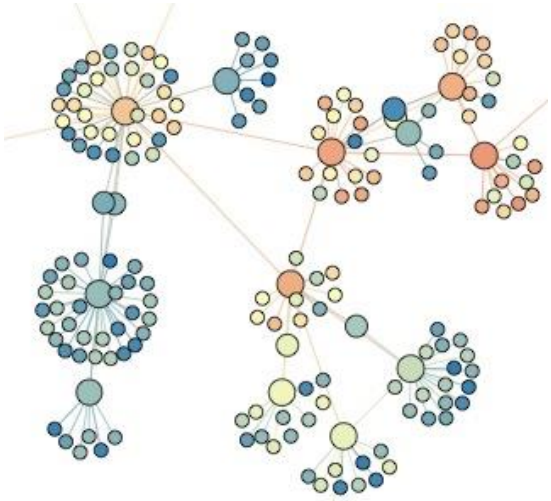


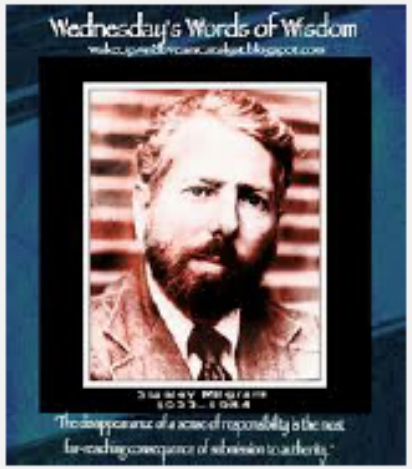
The small world phenomenon

Chapter 20

David Easley & Jon Klienberg

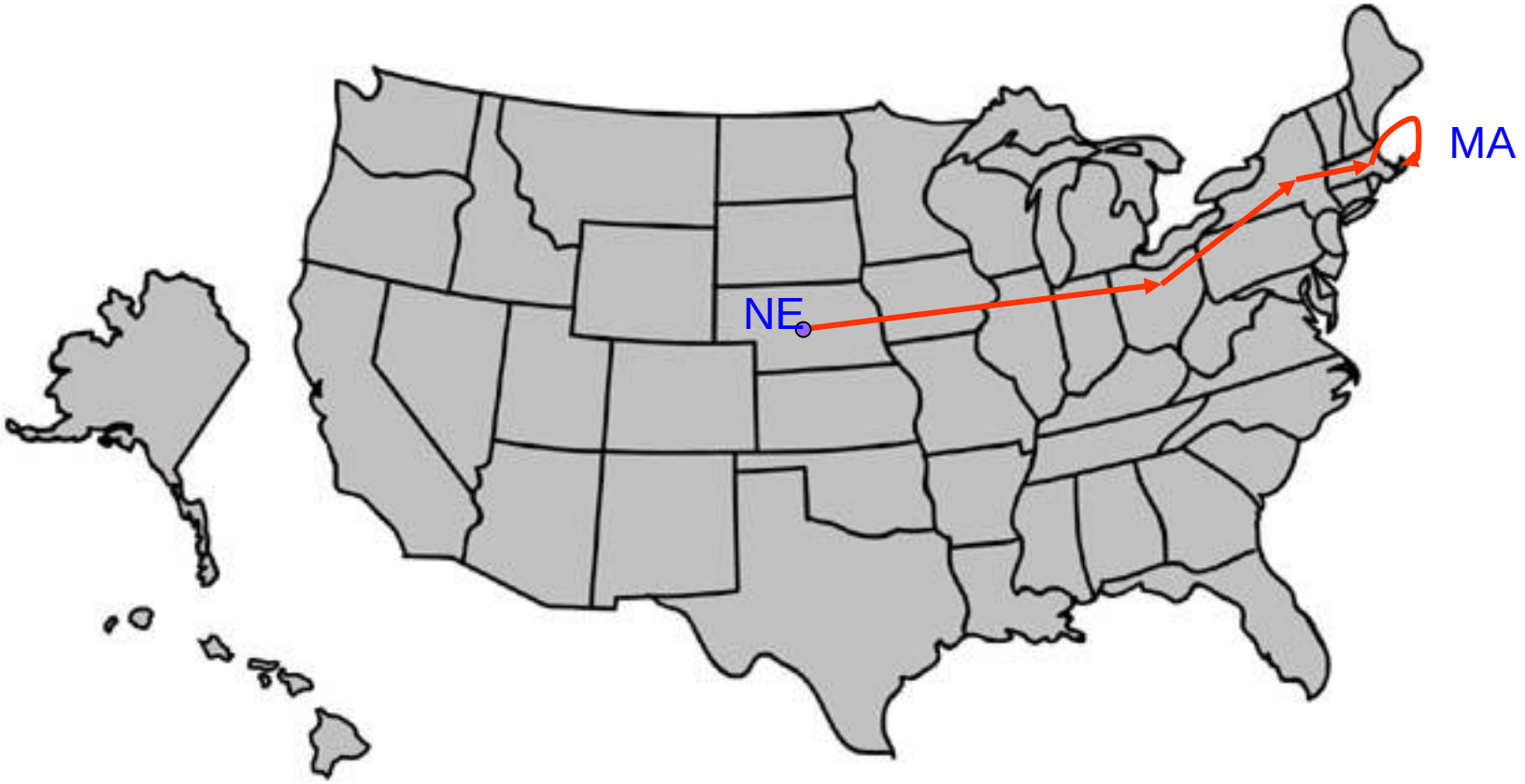
Social networks are rich in short paths, known as the *small-world phenomenon*, or the “six degrees of separation”

Milgram's experiment (Travers and Milgram, 1969)



http://stanleymilgram.com/milgram.php

Small world phenomenon:
Milgram's experiment



The first significant empirical study of the small-world phenomenon was undertaken by the social psychologist Stanley Milgram on the global friendship network as follows.

- Randomly chosen "starter" individuals each tries forwarding a letter to a designated "target" person living in the town of Sharon, MA, a suburb of Boston.
- The target's name, address, occupation, and some personal information are provided
- The participants could not mail the letter directly to the target; rather, each participant could only advance the letter by forwarding it to a single acquaintance that he or she knew on a first-name basis, with the goal of reaching the target as rapidly as possible.

Milgram's experiment results

Outcome:

20% of initiated chains reached target
average chain length = 6.5

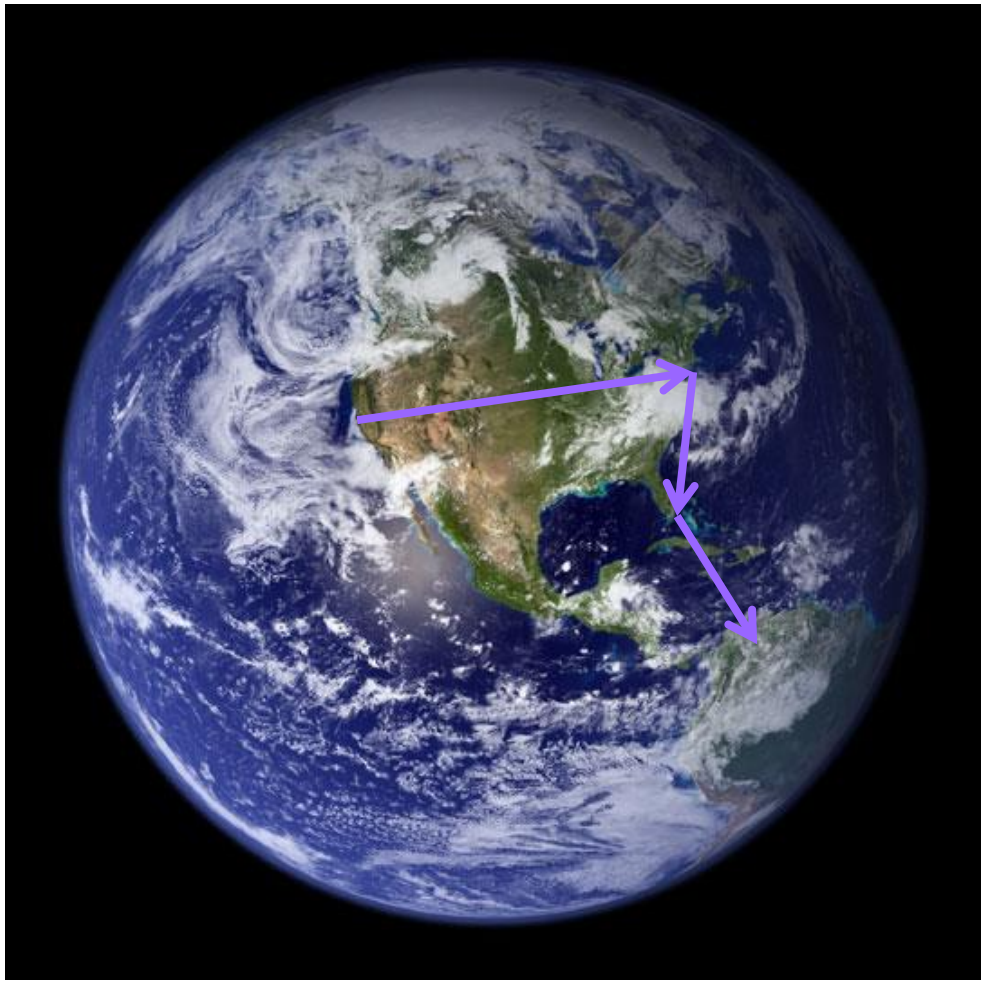
- "Six degrees of separation"

Milgram's experiment repeated

email experiment
Dodds, Muhamad, Watts,
Science 301, (2003)

- 18 targets
- 13 different countries

- 60,000+ participants
- 24,163 message chains
- 384 reached their targets
- average path length 4.0



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

Milgram's experiment really demonstrated two striking facts about large social networks:

1. short paths are abundant;

2. people, acting without any sort of global "map" of the network, are effective at collectively finding these short path.

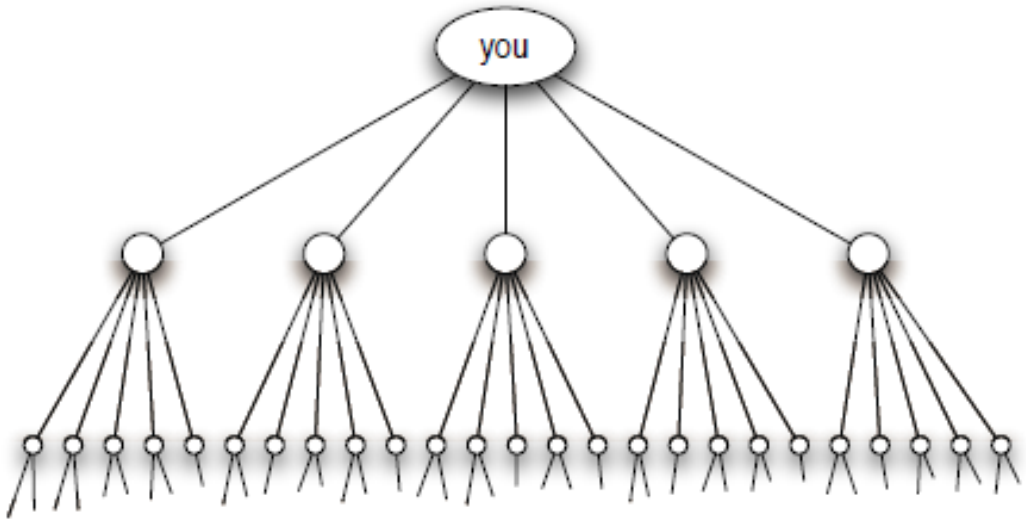
Structure and randomness
How people are able to find Short paths?

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

The Existence of Short path

Social networks expand to reach many people in only a few steps

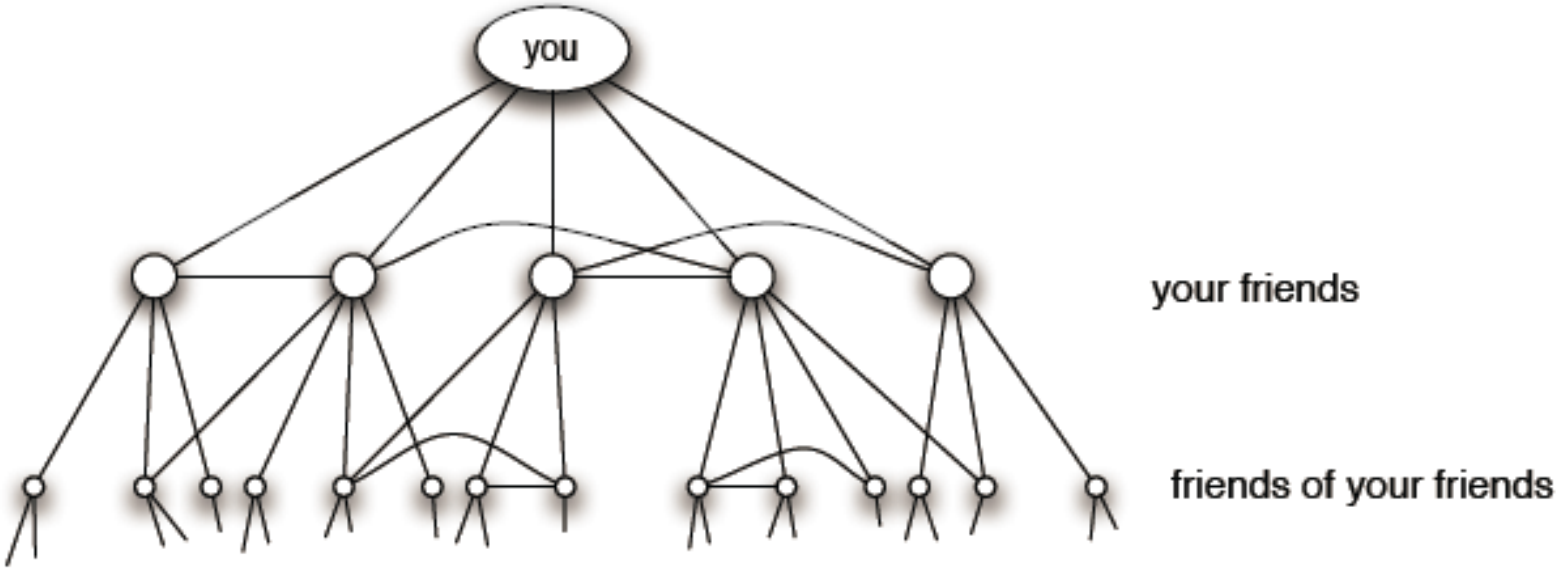


your friends 100

friends of your friends 100 x 100?

(a) *Pure exponential growth produces a small world*

10 billion after 5 steps!!!!



(b) *Triadic closure reduces the growth rate*

Network grows exponentially, leading to the existence of short paths!

The average person has between 500 and 1500 acquaintances, leading to $500^2 = 25K$ in one step, $500^3 = 125M$ in two steps, $500^4 = 62.5B$ in four (Figure (a)).

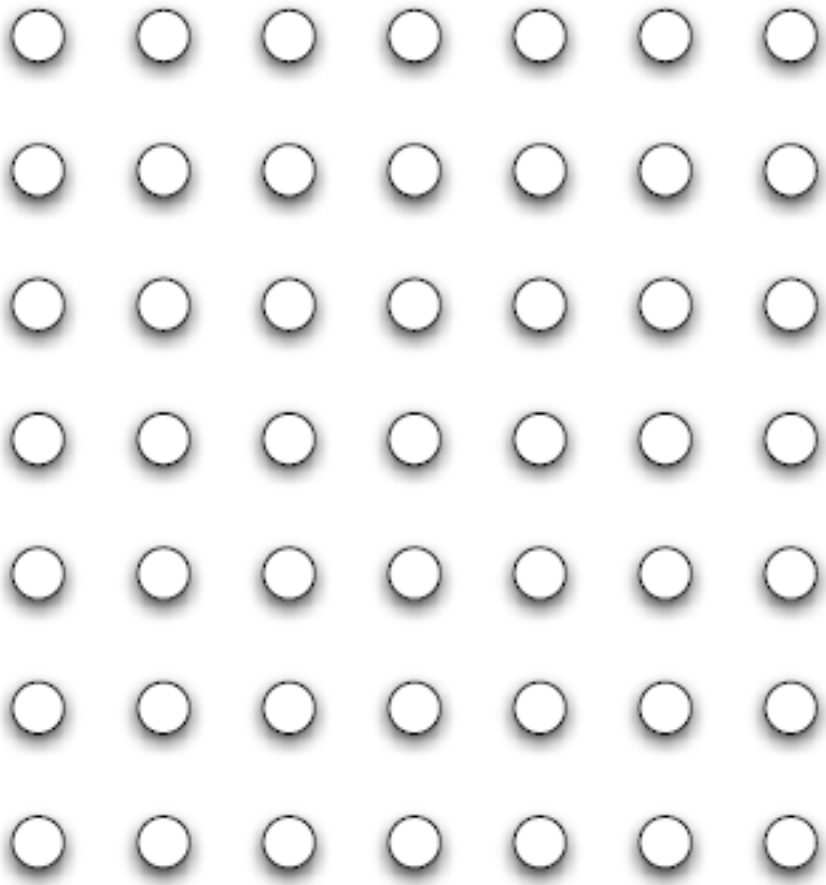
However, the effect of *triadic closure* works to limit the number of people one can reach by following short paths (Figure (b)).

Triadic closure: If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.

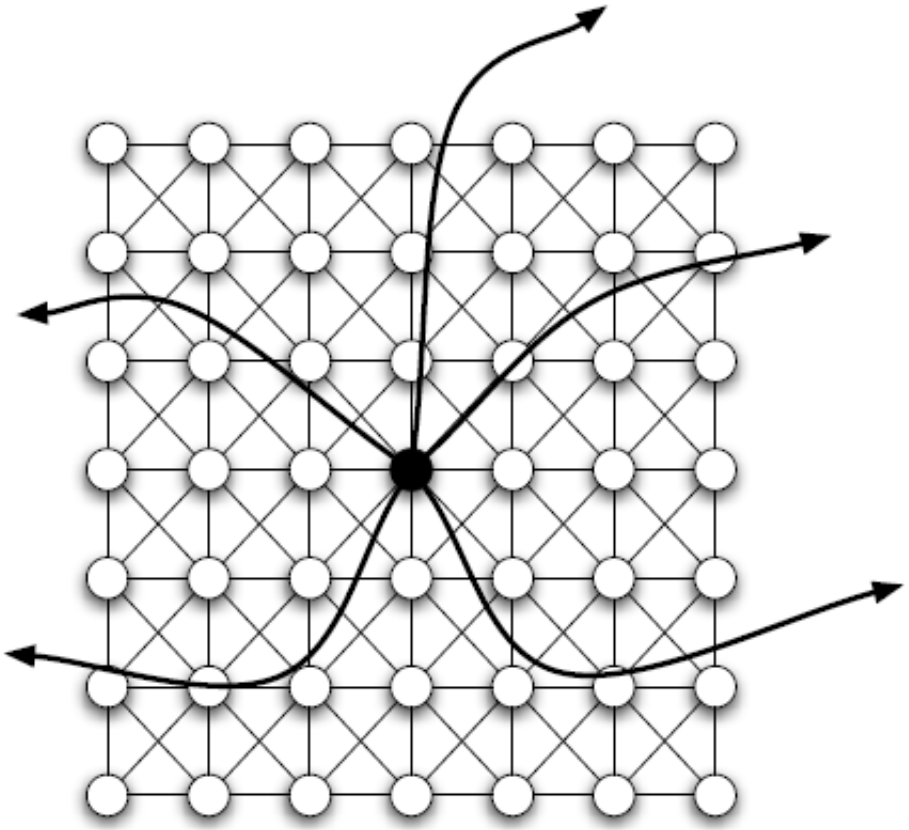
The Watts-Strogatz small-world network (1998)

- model follows a combination of two basic social-network ideas:

- Homophily: the principle that we connect to others who are like ourselves, and hence creates many triangles.
- Weak ties: the links to acquaintances that connect us to parts of the network that would otherwise be far away, and hence the kind of widely branching structure that reaches many nodes in a few steps.



(a) *Nodes arranged in a grid*

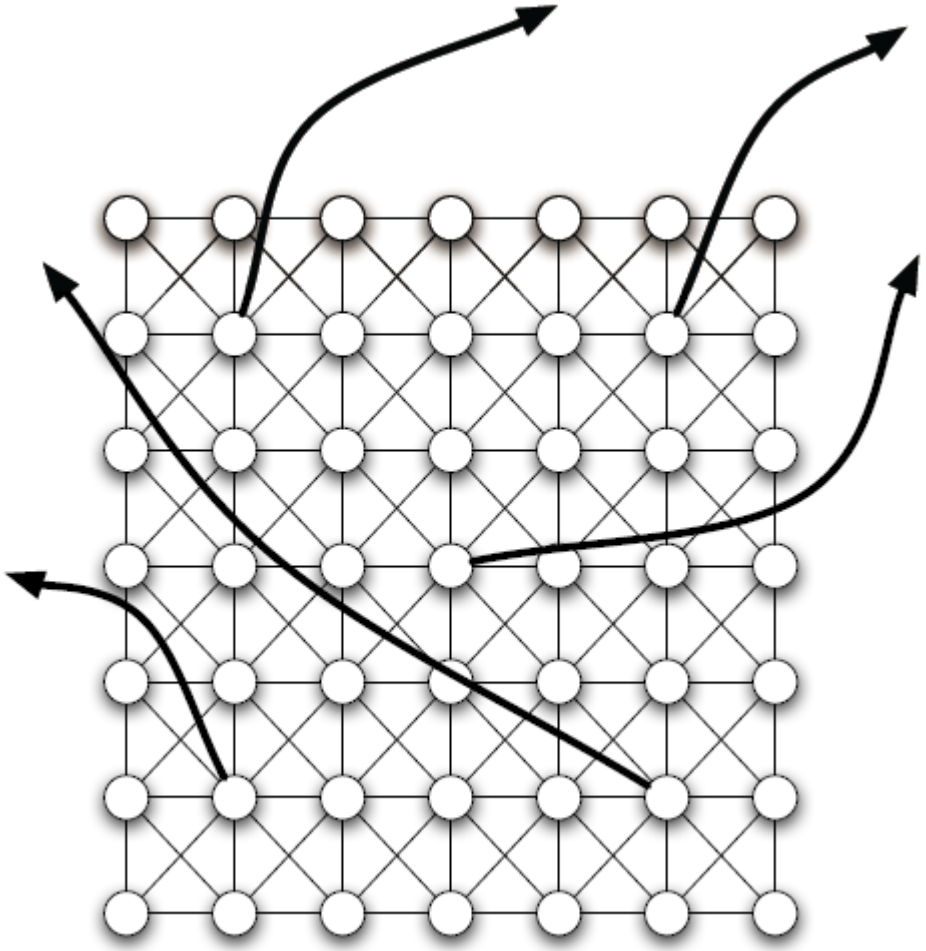


(b) *A network built from local structure and random edges*

- **Homophily – radius r**
- **Short paths - Link to k other nodes selected uniformly at random from the grid**

Suppose, for example, that instead of allowing each node to have k random friends, we only allow one out of every k nodes to have a single random friend - keeping the proximity based edges as before

The general conclusions of the Watts-Strogatz model still follow even if only a small fraction of the nodes on the grid each have a single random link.



The crux of the Watts-Strogatz model: introducing a tiny amount of randomness–in the form of long-range weak ties–is enough to make the world “small” with short paths between every pair of nodes.

Algorithmic issues

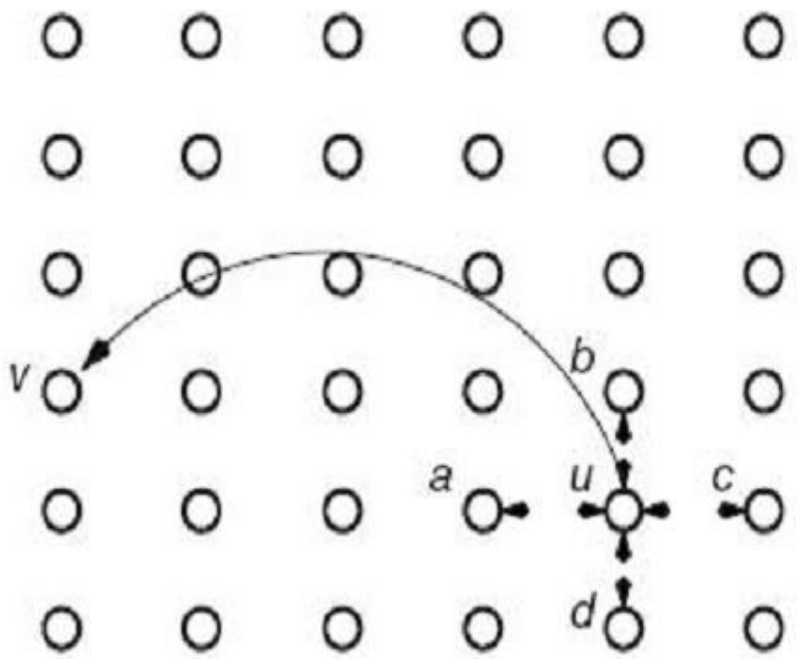
- The short path exists. (Structurally)
- But how can we find it?
- If we want the shortest path
  - "flooding" of the network (Instruct the starter to forward a letter to ***all*** of his friends, who in turn should have forwarded the letter to all of their friends, and so forth. This "flooding" of network would have reached the target as rapidly as possible. – bfs procedure, not feasible)
- Milgram's experiment
  - "tunneling" through the network (letter advancing just one person at a time – process that could have failed to reach target, even if a short path existed)
  - how to make decentralized routing so effective?

can we construct a random network in which decentralized routing succeeds, and if so, what are the qualitative properties that are crucial for success?

Decentralized Search Algorithm

- starting node s
- target node t
- seek to pass a message from s to t, by advancing the message along edges
- In each step, the current message holder v has knowledge of:
  - the underlying grid structure
  - the location of the target t on the grid ”
  - its own long-range contact
- The short path is unknown!

Short range and long range



In order to reach a far-away target, one must use long-range weak ties in a fairly structured, methodical way, constantly reducing the distance to the target.

Decentralized Search Algorithm (cont.)

- Delivery time: expected number of step required to reach the target

- The delivery time of any decentralized algorithm in the grid-based model is $\Omega(n^{2/3})$

  (Kleinberg, J., The small-world phenomenon: An algorithmic perspective. Proceedings of the 32nd Annual Symposium on Theory of Computing (2000))

The WS network is effective at capturing the density of triangles and the existence of short paths, but not the ability of people, working together in the network, to actually find the paths.

Weak ties that make the world small are too random in the WS model, since , they are completely unrelated to the similarity among nodes that produces homophily-based links, they are hard for the people to use reliably!

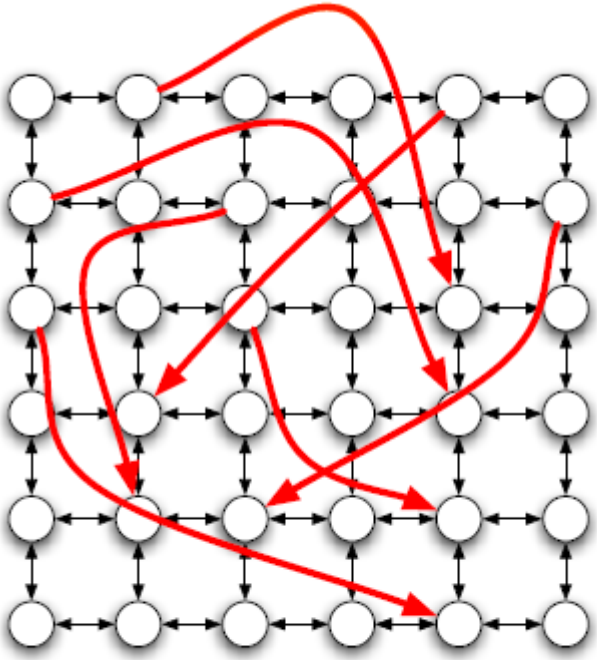
Generalizing the network model

- Introducing parameter q >= 0 (clustering exponent)
- For two nodes v and w :

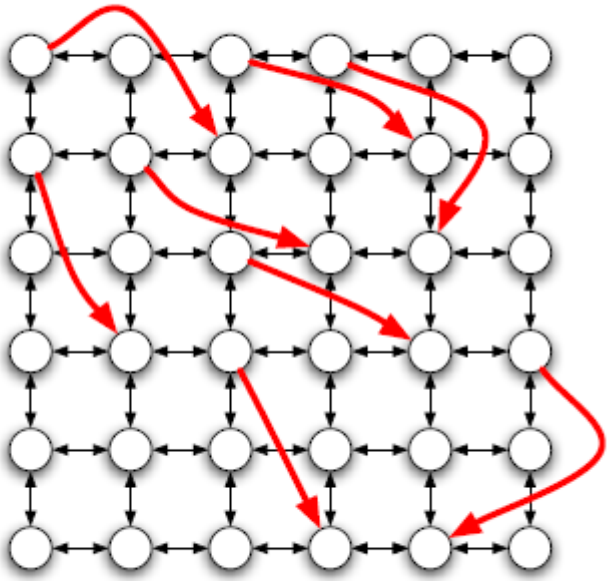
grid distance $d(v, w)$: the number of edges in a shortest path between them on the grid.

$d(v,w)^{-q}$: the probability to choose w as the long-range contact for v

- q too small : too random(uniformly) & cannot be used effectively for decentralized search
- q too large : not random enough; don’t provide enough long distance jumps that are needed to create a small world
- q=0; original grid based model; links chosen uniformly at random

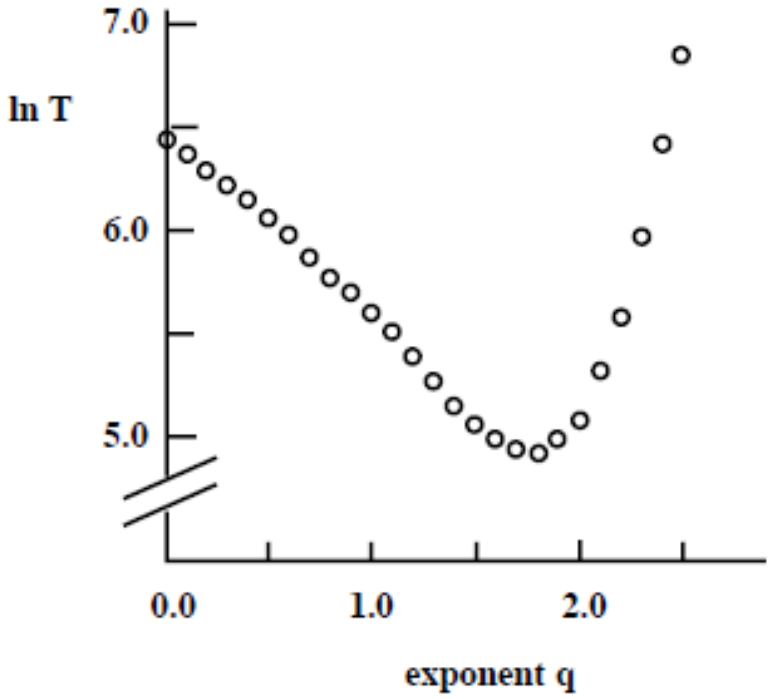


(a) *A small clustering exponent*



(b) *A large clustering exponent*

With a small clustering exponent, the random edges tend to span long distances on the grid; as the clustering exponent increases, the random edges become shorter.



Simulation of decentralized search in the grid-based model with clustering exponent q. Each point is the average of 1000 runs on (a slight variant of) a grid with 400 million nodes. The delivery time is best in the vicinity of exponent q = 2, as expected; but even with this number of nodes, the delivery time is comparable over the range between 1:5 and 2

Decentralized search is most effective when q=2;

Random links follow an inverse square distribution

The probability that a random edge links into some node in this ring is approximately independent of the value of d.

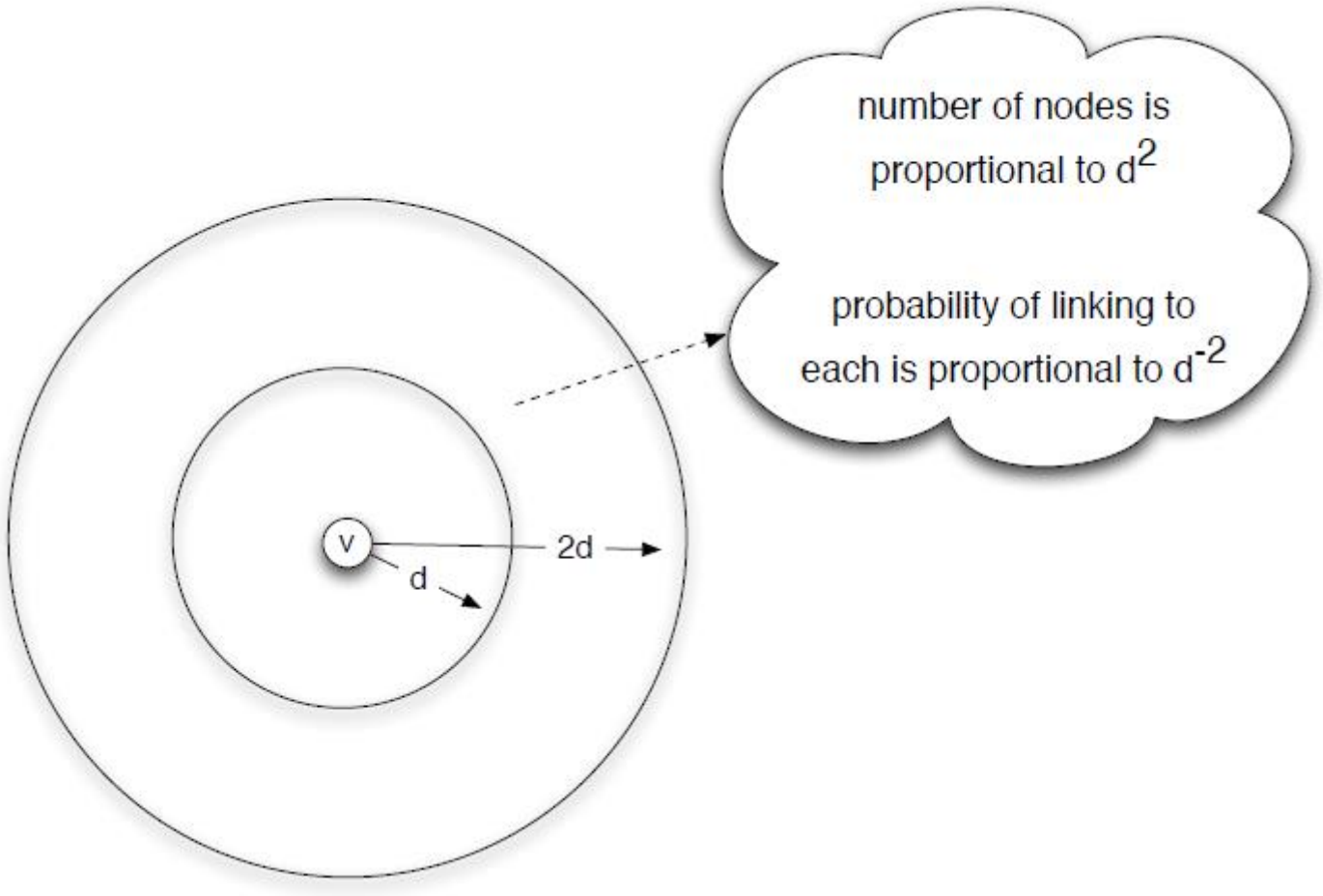
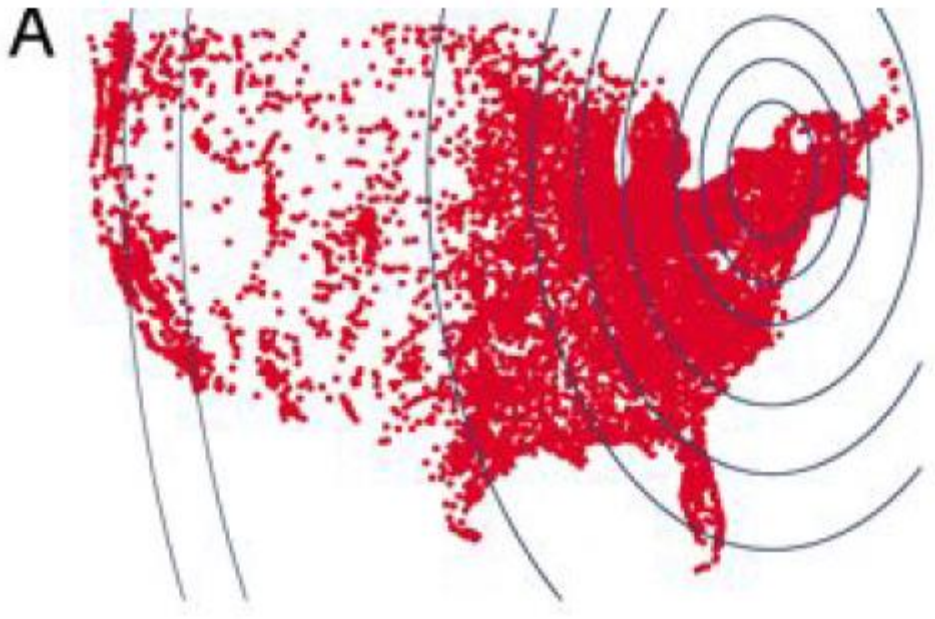


Fig. The concentric scales of resolution around a particular node

Geographic data on friendship



The population density of the LiveJournal network studied by Liben-Nowell et al.

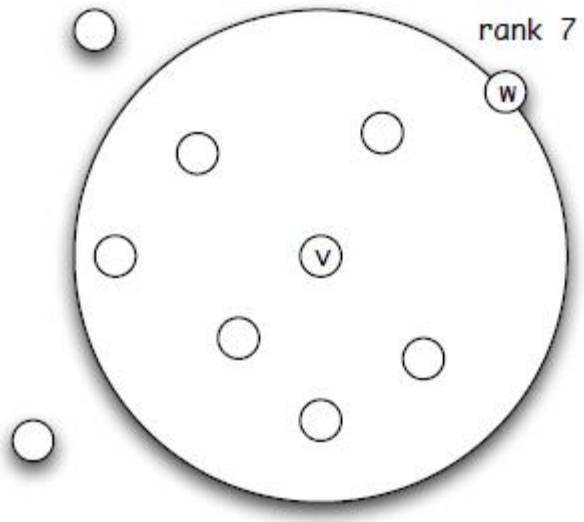
Challenge: distribution of nodes is typically not uniform

Possible solution: determine link probs not by distance but by **rank**
For v, **rank (w) = # nodes close to v than w**

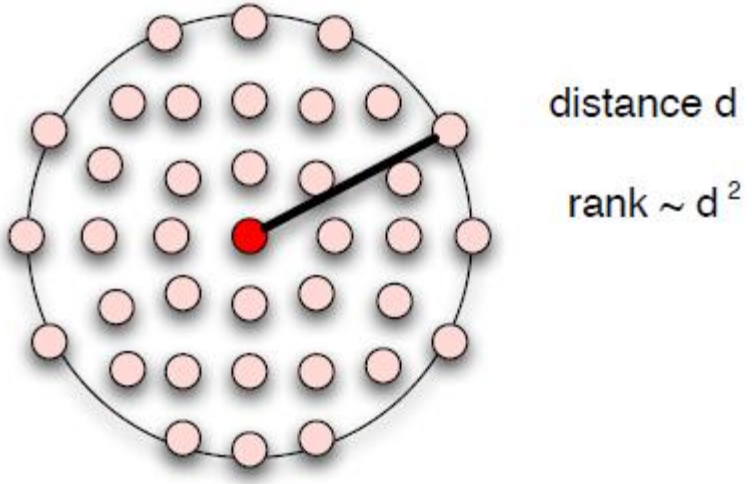
Rank-based friendship: **prob (link v-w) = rank(w)$^{-p}$.**

For uniformly distributed nodes: linking with prob d^{-2} roughly corresponds to linking with prob rank(w)$^{-1}$.

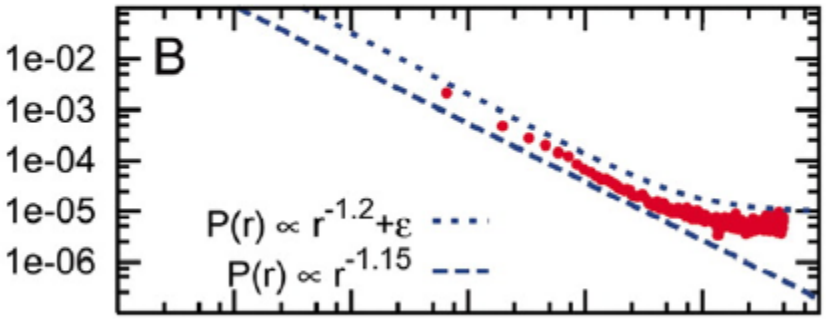
More generally: **p=1 yields efficient decentralized search.** (Liben-Nowell et al. 2005)
(some empirical support that this is how people link)



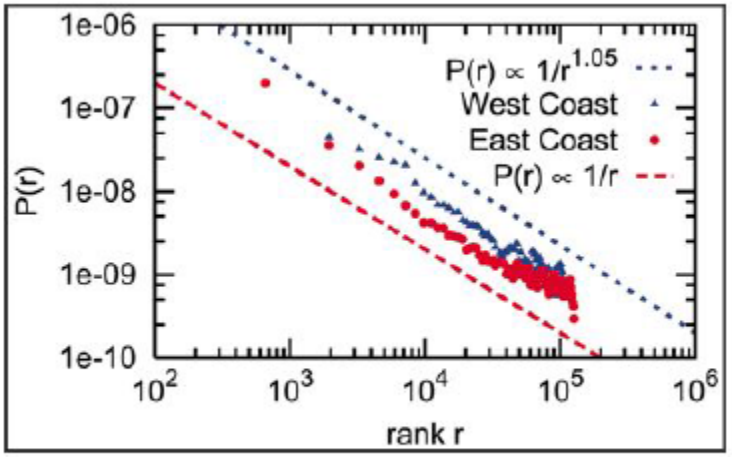
(a) *w is the 7th closest node to v.*



(b) *Rank-based friendship with uniform population density.*



(a) *Rank-based friendship on LiveJournal*



(b) *Rank-based friendship: East and West coasts*

The probability of a friendship as a function of geographic rank on the blogging site LiveJournal.

Recall: social focus = any type of community, occupaation, etc., that serves to organize social life.

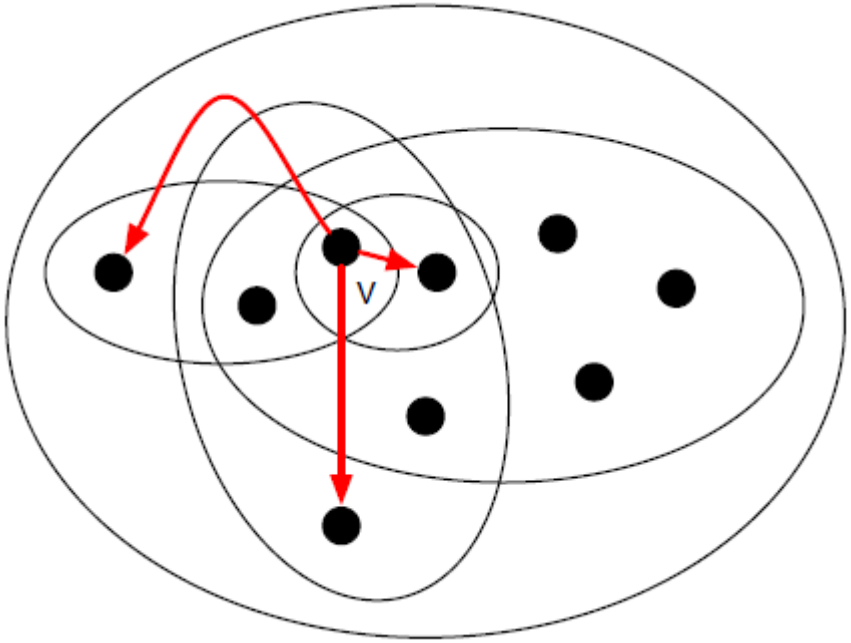
Social Distance between v and w (dist(v,w)): smallest social focus that includes both.

Make link v-w with prob $dist(v,w)^{-p}$.

Again: $p=1$ yields efficient decentralized search (Kleinberg 2001).

Bottom line: model based on uniformly distributed nodes and geographical distance can easily be extended: efficient search is still possible.

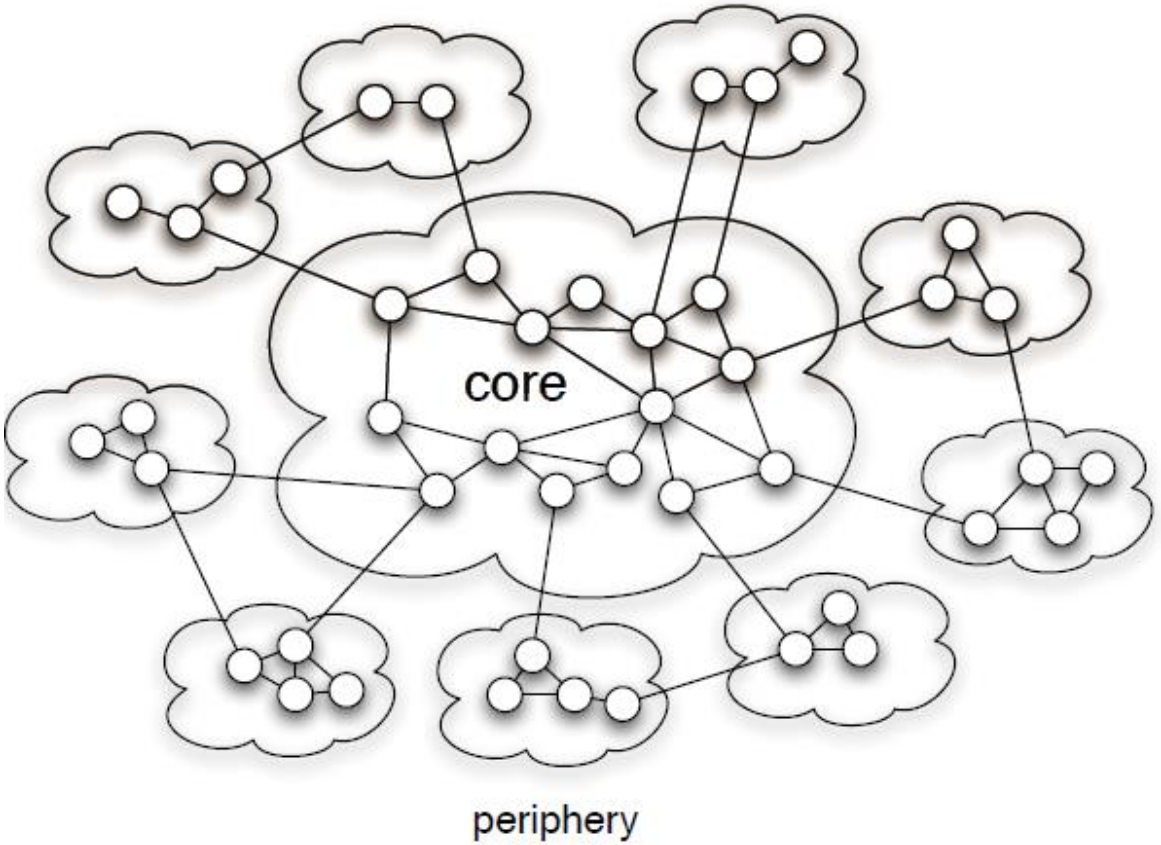
Social Foci and Social distance



When nodes belong to multiple foci, we can define the social distance between two nodes to be the smallest focus that contains both of them. In the figure, the foci are represented by ovals; the node labeled v belongs to five foci of sizes 2; 3; 5; 7, and 9 (with the largest focus containing all the nodes shown).

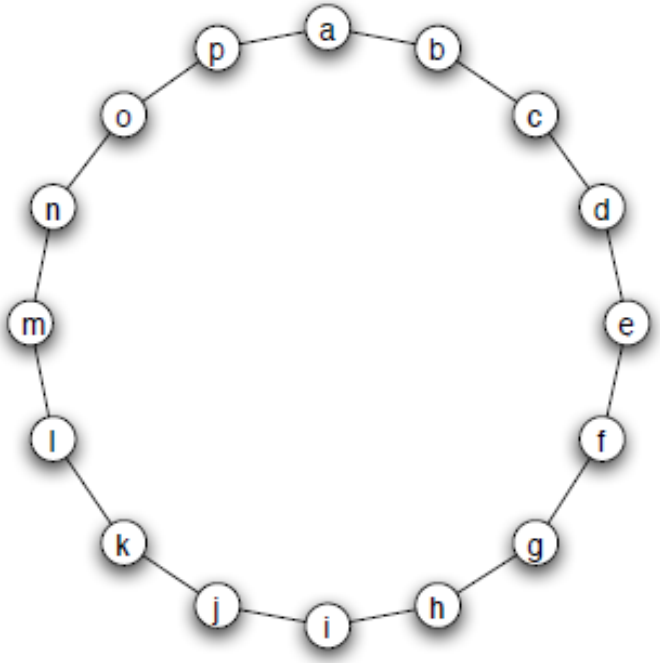
the success rate at finding targets in recreations of the Milgram experiment has often been much lower than it was in the original work. {Judith Kleinfeld]

The core-periphery structure of social networks & difficulties in decentralized search

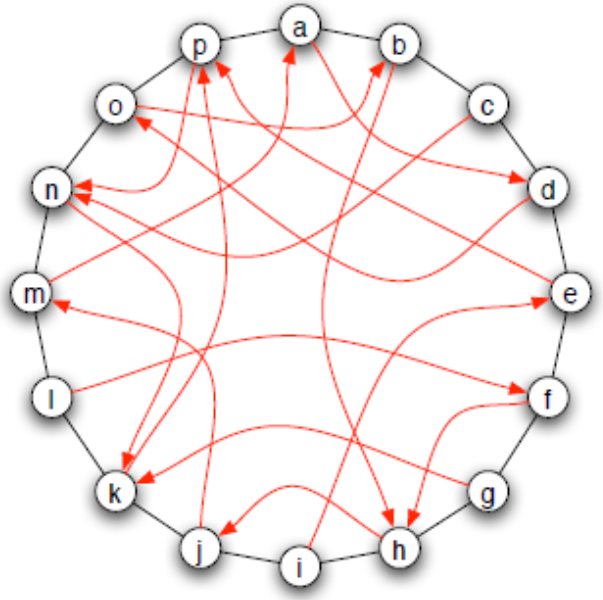


it is harder for Milgram style decentralized search to find low-status targets than high-status targets

Analysis of decentralized search: (A) optimal exponent in one dimension



(a) *A set of nodes arranged in a ring.*



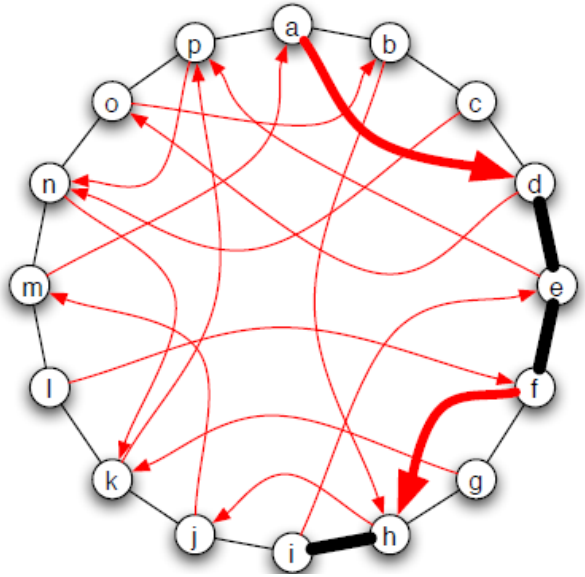
(b) *A ring augmented with random long-range links.*

Local contacts: adjacent edges of ‘v’
Contact: node to which ‘v’ has an edge
Long-range contact: other remaining nodes
One-dimensional version of grid with random edges

Analysis of decentralized search: (B) Myopic search (q=1)

- Choose random s and t
- Each intermediate node knows location c only its neighbors and t
- Full network is not known to intermediat nodes
- Example: s=a and t=i
- Path: a-d-e-f-h-i : five steps path

- NOTE: myopic path from a to i is not shortest (a-b-h-i)
  - Lack of knowledge about full network structure
  - Myopic search finds paths that are surprisingly short!



Problem at hand:

- **generate random network with such weak ties**
- **choose random starting and target nodes**
- **number of steps needed by myopic search is X**

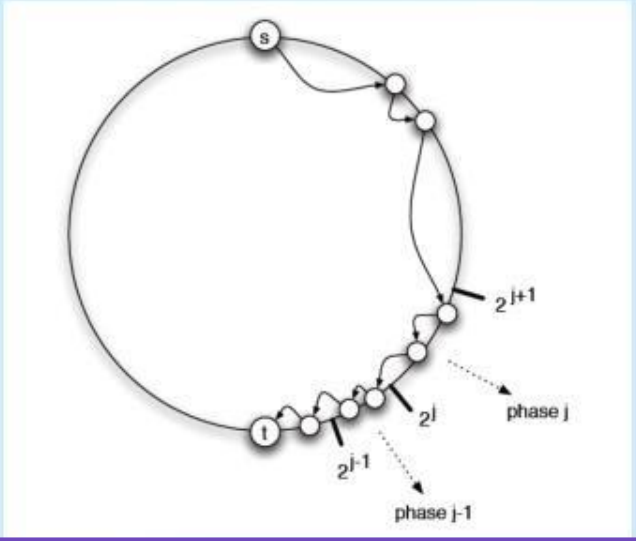
⇒ Show that E(X) is 'small'

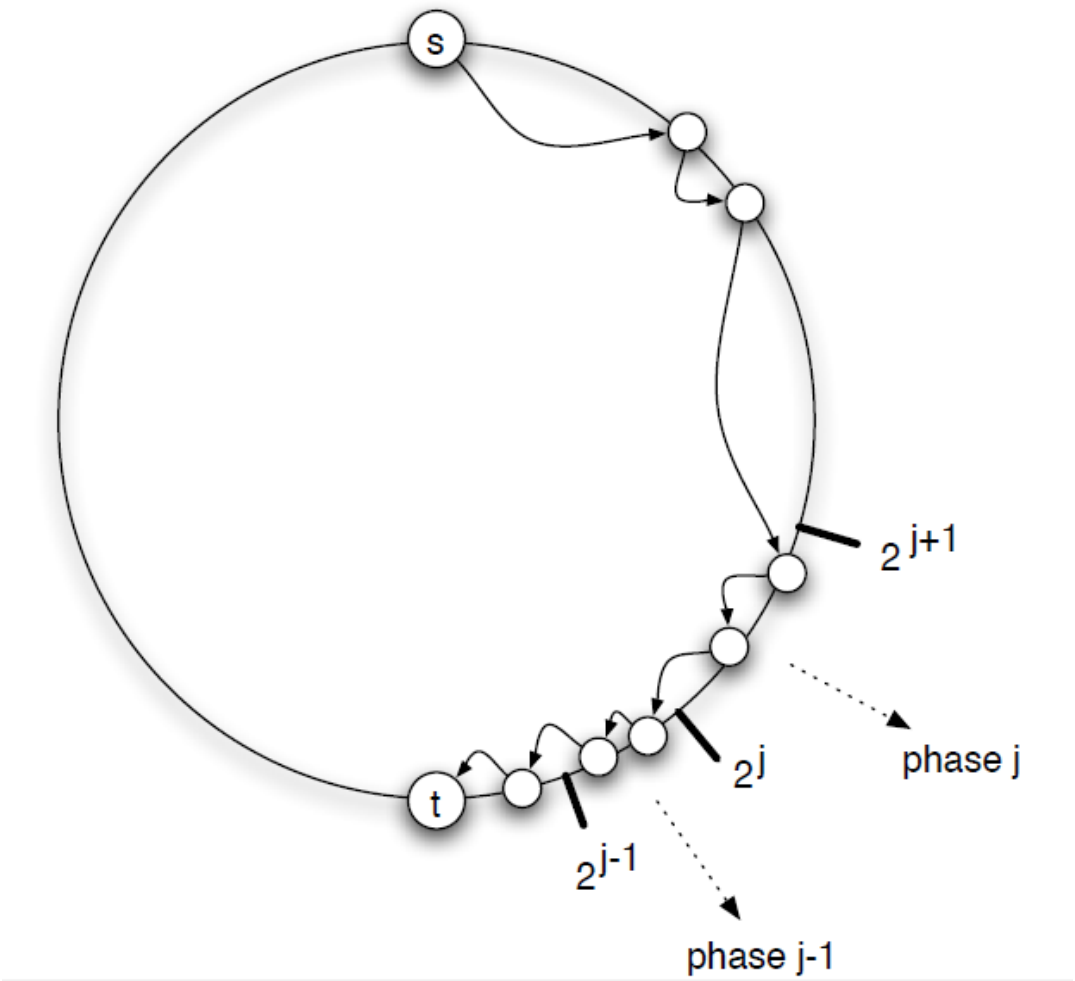
strategy: track # steps needed to halve distance

Define: message is in phase j if its distance from target is between 2^j and 2^{j+1}

The number of phases is at most $\log_2 n$ (from $2^j = n$)

$$E(X) = E(X_1) + \ldots + E(X_{\log_2 n})$$





$$X = X_1 + X_2 + \quad + X_{\log n};$$

Linearity of expectation - the expectation of a sum of random variables is equal to the sum of their individual expectations

$$E[X] = E[X_1 + X_2 + \quad + X_{\log n}] = E[X_1] + E[X_2] + \quad + E[X_{\log n}]$$

$E[X]$ is at most proportional to $(\log n)^2$.

The Normalizing Constant

the probability of v linking to w =1/Z $d(v,w)^{-1}$

there are two nodes at distance 1 from v, two at distance 2,… two at each distance d up to n/2.

Assuming n is even, there is also a single node at distance n/2 from v ; the node diametrically opposite it on the ring.

$$Z \leq 2\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n/2}\right)$$

$$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{k} \leq 1+\int_1^k \frac{1}{x}dx = 1+\ln k.$$

k = n/2

$$Z \leq 2(1+\ln(n/2)) = 2+2\ln(n/2).$$

w.k.t. $\ln x \leq \log_2 x$

$$Z \leq 2 + 2\log_2(n/2) = 2 + 2(\log_2 n) - 2(\log_2 2) = 2\log_2 n.$$

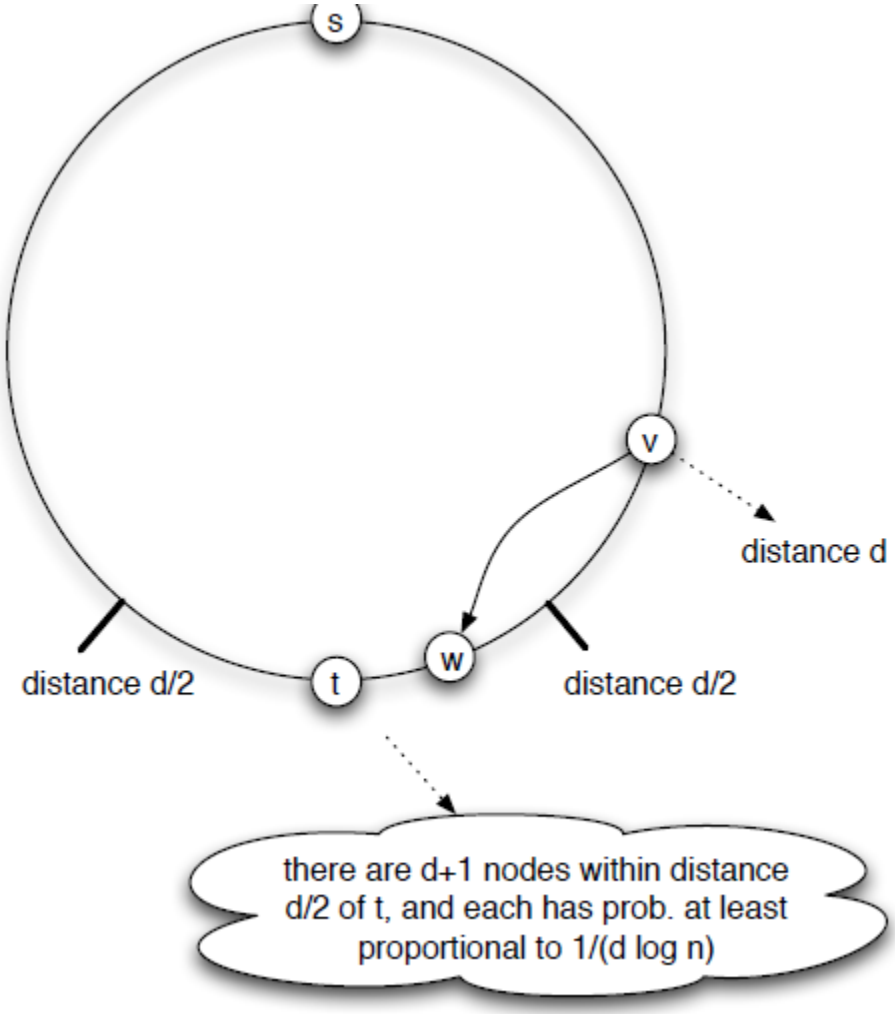
the probability v links to w is

$$\frac{1}{Z} d(v,w)^{-1} \geq \frac{1}{2\log n} d(v,w)^{-1}$$

Analyzing the Time Spent in One Phase of Myopic Search

when the message is at a node v whose distance to the target t is some number d between 2^j and 2^{j+1}

One way for the phase to come to an end immediately would be for v's long-range contact w to be at distance <= d/2 from t.



with reasonable probability, v's long-range contact lies within half the distance to the target.

d/2 nodes consecutively on each side of a node and a long range

$$\frac{1}{2\log n} d(v,w)^{-1} \geq \frac{1}{2\log n} \cdot \frac{1}{3d/2} = \frac{1}{3d\log n}$$

the probability that one of the nodes is the long-range contact of v is at least

$$d \cdot \frac{1}{3d\log n} = \frac{1}{3\log n}$$

If one of these nodes is the long-range contact of v, then phase j ends immediately in this step.

.
.
.

$$E[X_j] \leq 3\log n.$$

E [X] is a sum of the log n terms $E[X_1]+E[X_2]+ \ +E[X_{\log n}]$

$$E[X] \leq 3(\log n)^2.$$

Myopic search constructs a path that is *exponentially smaller* than # nodes