Online Social Networks and Media

Navigation in a small world

Small world phenomena

- Small worlds: networks with short paths

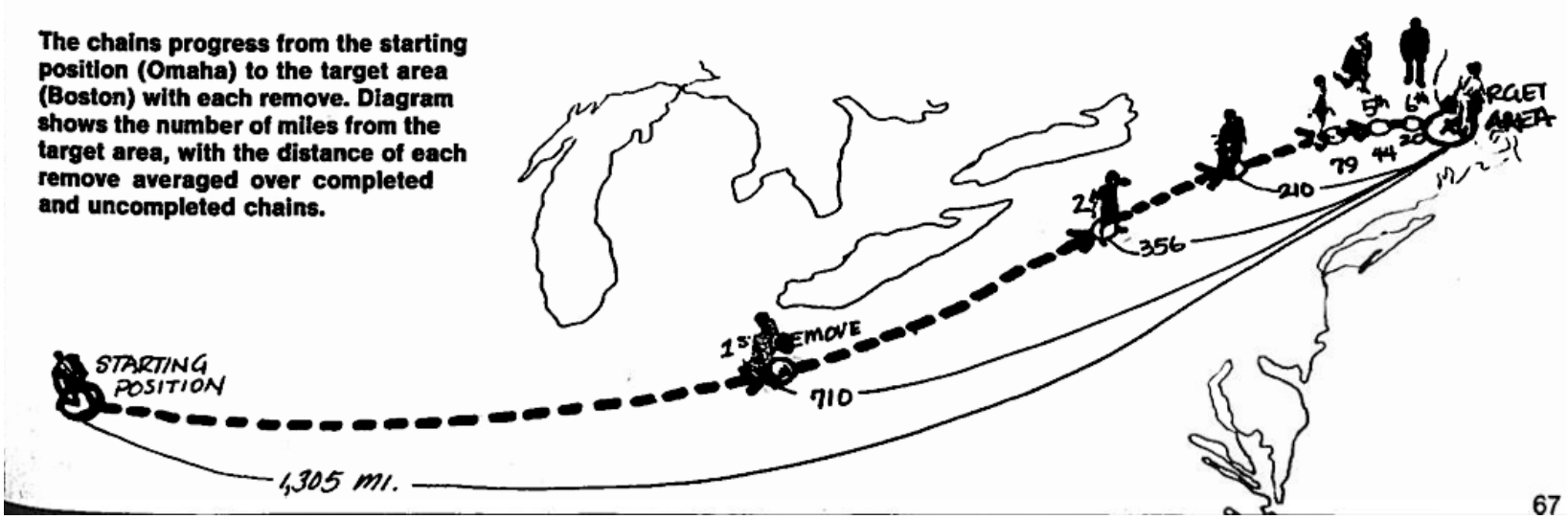
Stanley Milgram (1933-1984): “The man who shocked the world”

Obedience to authority (1963)

Small world experiment (1967)

Small world experiment

- Letters were handed out to people in Nebraska to be sent to a target in Boston
- People were instructed to pass on the letters to someone they knew on first-name basis
- The letters that reached the destination followed paths of length around 6
- Six degrees of separation: (play of John Guare)

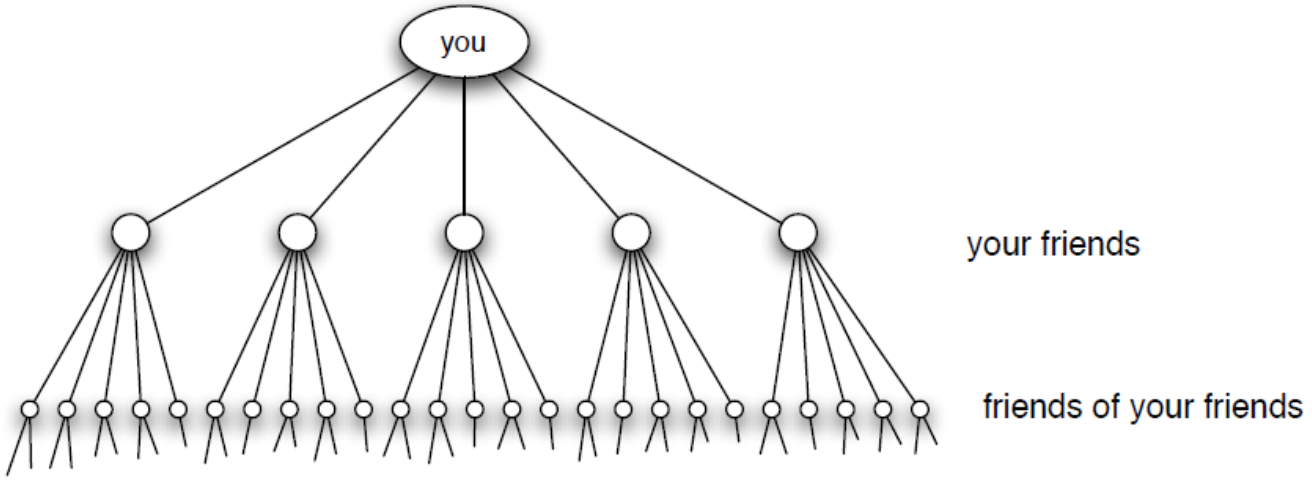


Milgram's experiment revisited

- What did Milgram's experiment show?
  - (a) There are short paths in large networks that connect individuals
  - (b) People are able to find these short paths using a simple, greedy, decentralized algorithm

A simple explanation

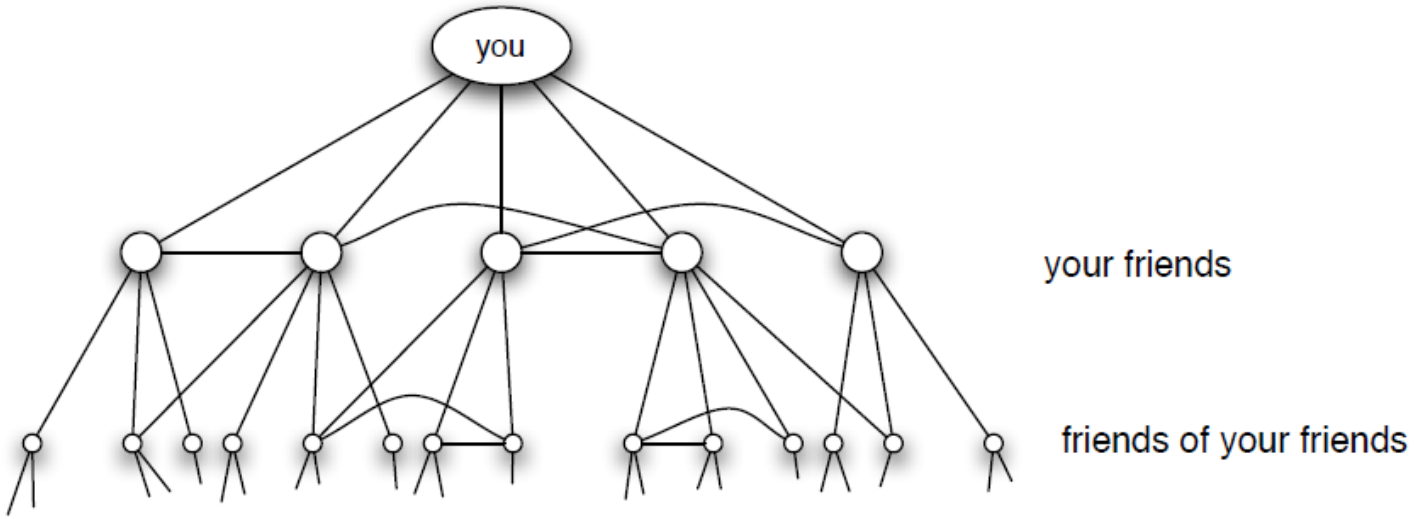
- If everyone had different friends, number of people reachable increases exponentially with the size of the path



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

Simple explanation does not work

- As we have seen friendships tend to form triangles, so the previous assumption is not true.

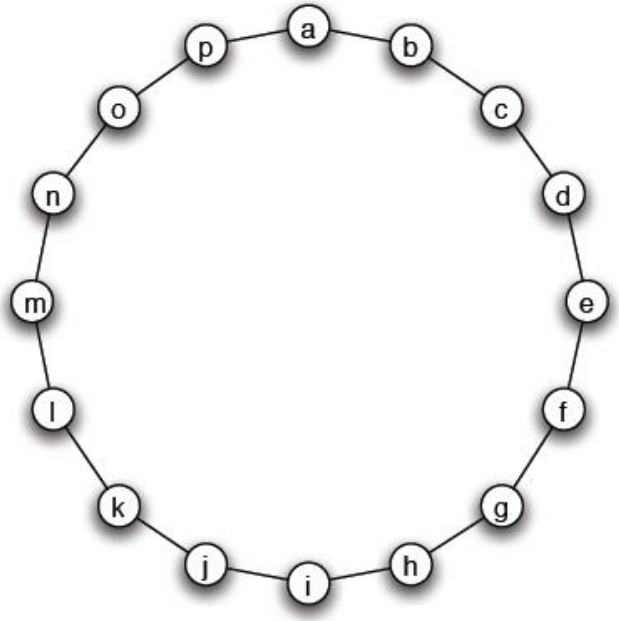


(b) *Triadic closure reduces the growth rate*

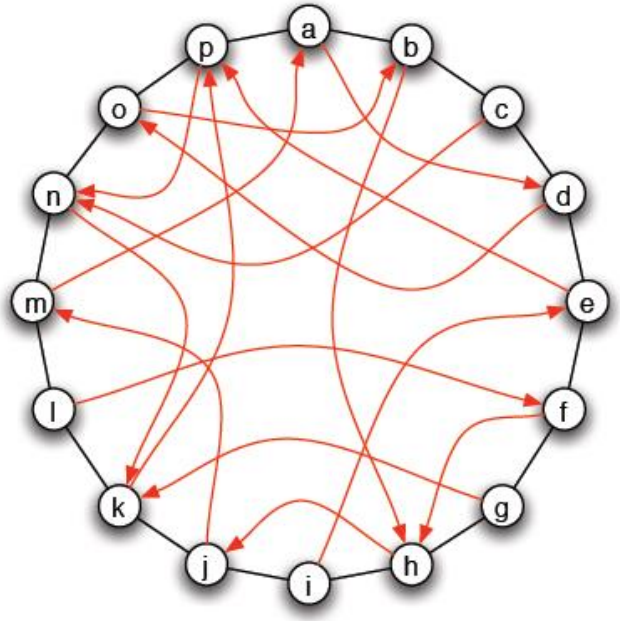
- How can we have both many triangles and short paths?

Small worlds

- We can construct graphs with short paths
  - E.g., the Watts-Strogatz model



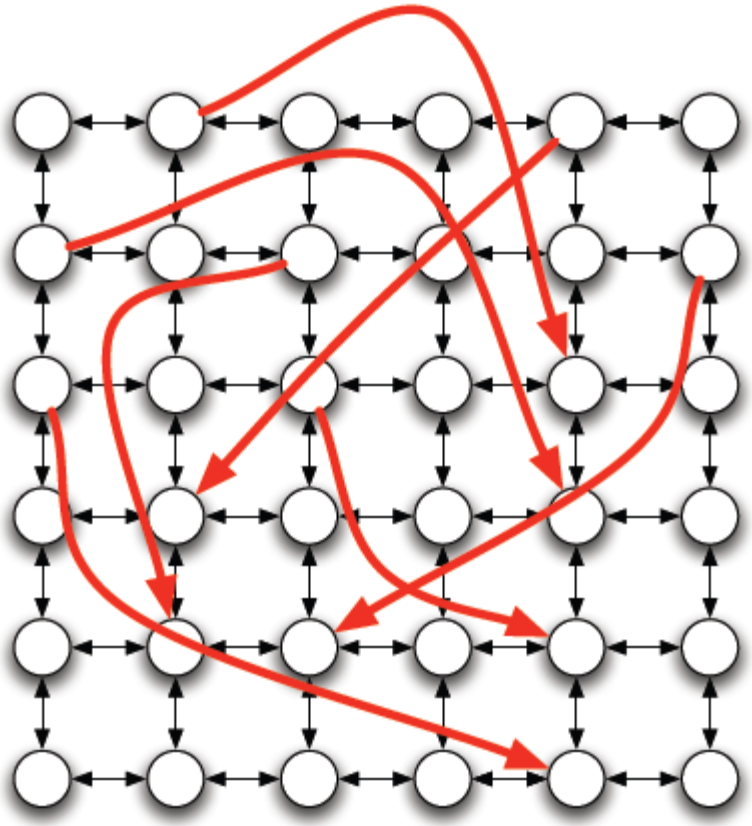
(a) *A set of nodes arranged in a ring.*



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

Small worlds

- Same idea to different graphs



Explanation

- Assume that we add k random links from every node.
- Looking at the graph formed by the random links it is unlikely that we have many common neighbors
  - Therefore, we have almost exponential growth of the reachable nodes.

Navigation in a small world

- Kleinberg: Many random graphs contain short paths, but how can we find them in a decentralized way?
- In Milgram's experiment every recipient acted without knowledge of the global structure of the social graph, using only
  - information about geography
  - their own social connections

Kleinberg's navigation model

- Assume a graph similar (but not the same!) to that of Watts-Strogatz
  - There is some underlying "geography": ring, grid, hierarchy
    - Defines the local contacts of a node
    - Enables to navigate towards a node
  - There are also shortcuts added between nodes
    - The long-range contacts of a node
    - Similar to WS model – creates short paths

Kleinberg's navigational model

- Given a source node s, and a navigation target t we want to reach, we assume
  - No centralized coordination
    - Each node makes decisions on their own
  - Each node knows the "geography" of the graph
    - They can always move closer to the target node
  - Nodes make decisions based only on their own contacts (local and long-range)
    - They do not have access to other nodes' contacts
  - No flooding is allowed
    - A node cannot send the message to all of her friends.
  - Greedy (myopic) decisions
    - Always move to the node that is closest to the target.

Example

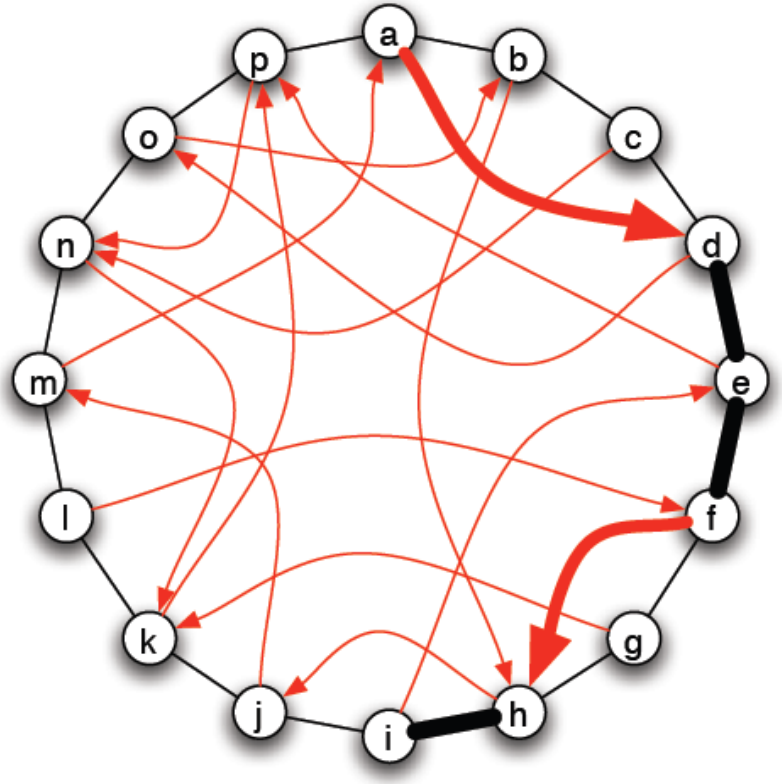


Figure 20.15: In myopic search, the current message-holder chooses the contact that lies closest to the target (as measured on the ring), and it forwards the message to this contact.

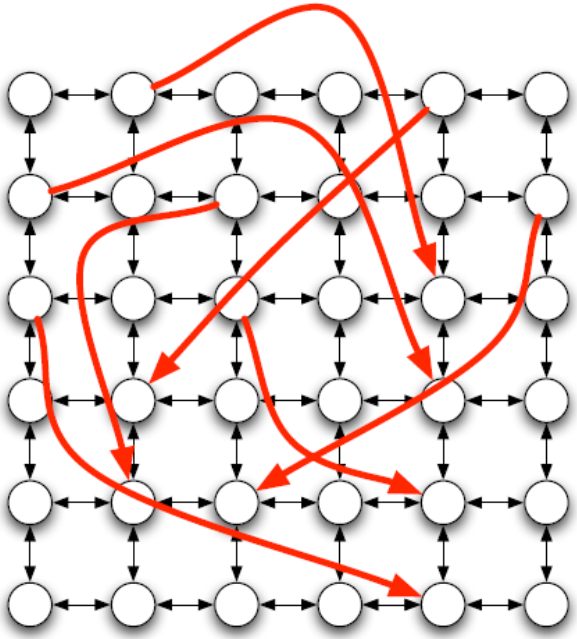
Long-range contacts

- If long-range contacts are created uniformly at random they do not help in navigation/search.
  - Proven theoretically
- We create contacts with probability that decreases with the distance to the endpoint

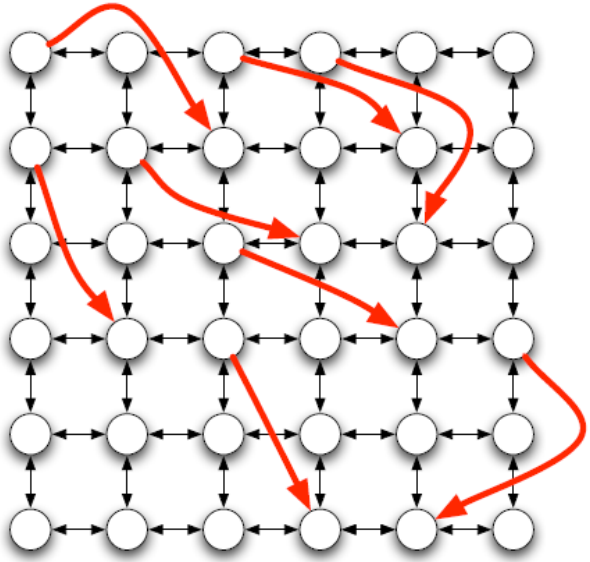
$$P(x \to y) \sim d(x,y)^{-q}$$

  - q: clustering exponent
  - When q = 0, uniform selection
  - When q > 0, nodes are more likely to connect closer to them (follows also intuitively)

Clustering exponent



(a) *A small clustering exponent*



(b) *A large clustering exponent*

Clustering exponent

- The right (clustering) exponent q depends on the geography:
  - q = 1 for a 1-dimensional ring, q = 2 for a 2-dimensional grid.
- This exponent is the only one for which greedy search follows "short" (polylogarithmic length) paths

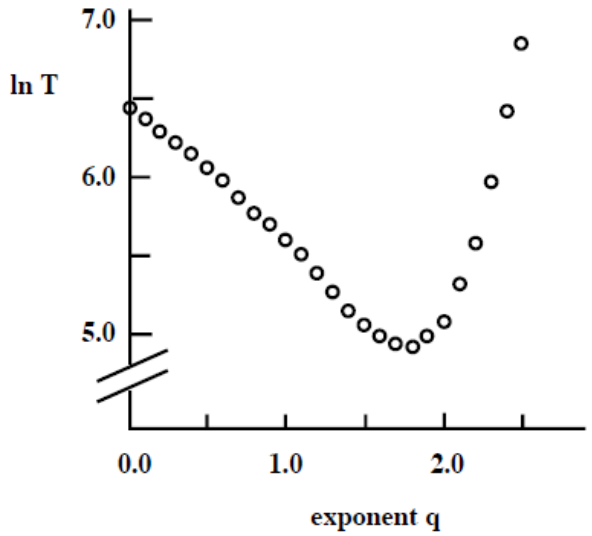


Figure 20.6: Simulation of decentralized search in the grid-based model with clustering exponent q. Each point is the average of 1000 runs on (a slight variant of) a grid with 400 million nodes. The delivery time is best in the vicinity of exponent $q = 2$, as expected; but even with this number of nodes, the delivery time is comparable over the range between 1.5 and 2 [248].

Theoretical results

- Proven for an underlying grid:
  - If the underlying topology is a 2-dimensional grid, and the clustering exponent is $q = 2$, then the search time is $O(\log^2 n)$. If $q \neq 2$, then the search time is $O(n^c)$ for some c > 0.

- Exact same theorem for q = 1 for the ring.
  - If the underlying topology is a 1-dimensional ring, and the clustering exponent is $q = 1$, then the search time is $O(\log^2 n)$. If $q \neq 1$, then the search time is $O(n^c)$ for some c > 0.

- Extends to any dimension d
  - We obtain $O(\log^2 n)$ search time when q=d, the exponent is equal to the dimension of the underlying graph

Proof Intuition

- The algorithm has the same probability to link to any scale of resolution

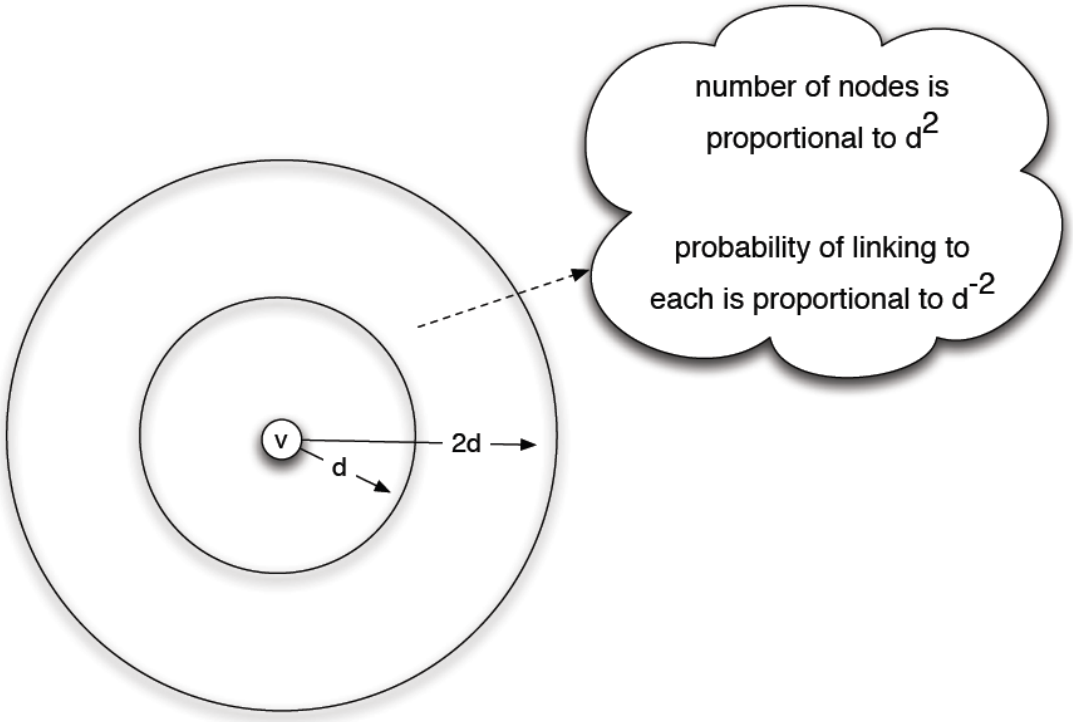
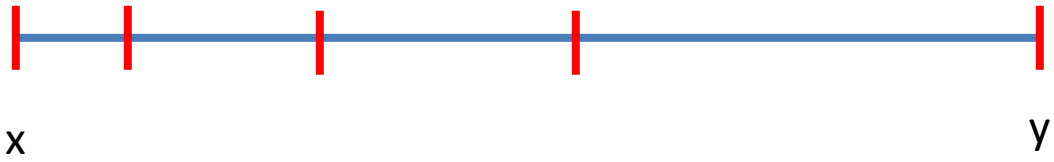


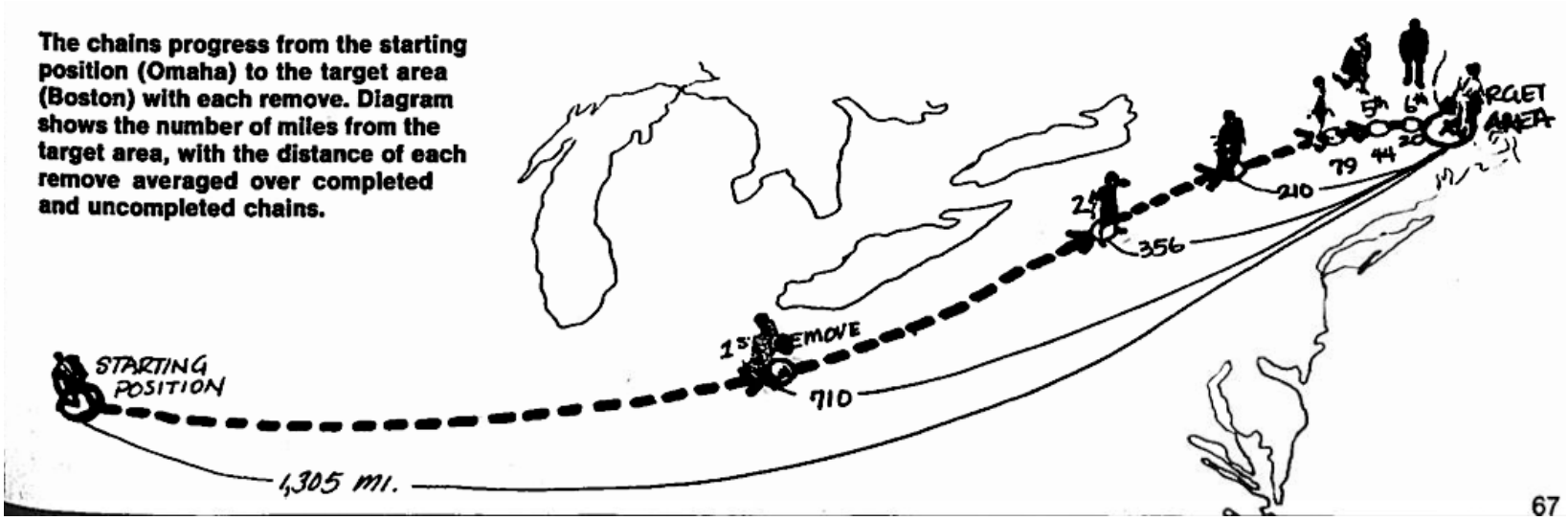
Figure 20.7: The concentric scales of resolution around a particular node.

Proof intuition

- The algorithm is able to replicate what happens in the Milgram experiment

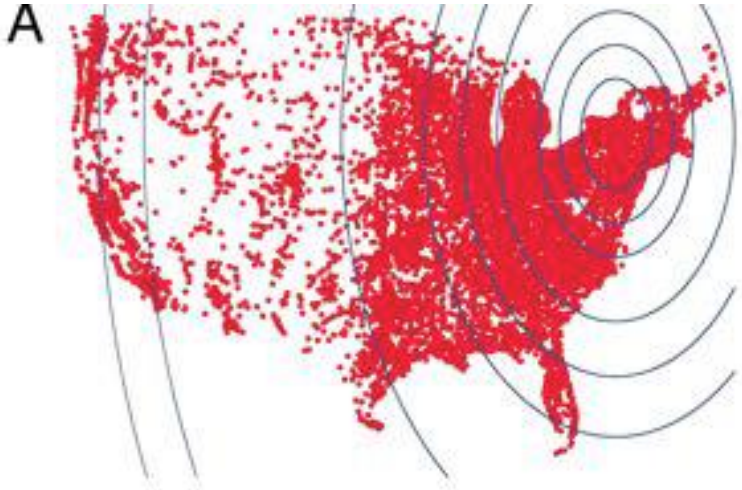


logn scales, logn steps in expectation to change scale



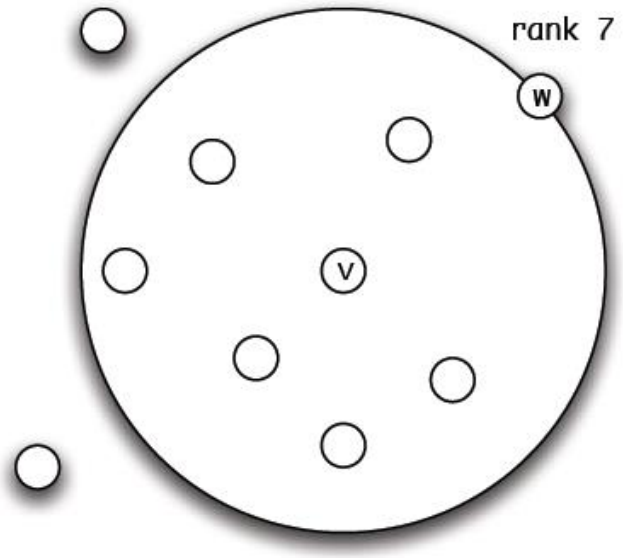
Long range links in the real world

- Is it the case that people link to each other with probability $P(x \to y) \sim d(x,y)^{-2}$?
- Live Journal data
  - Connections between friends
  - Postal codes for locations
- Problem: non-uniform density of points

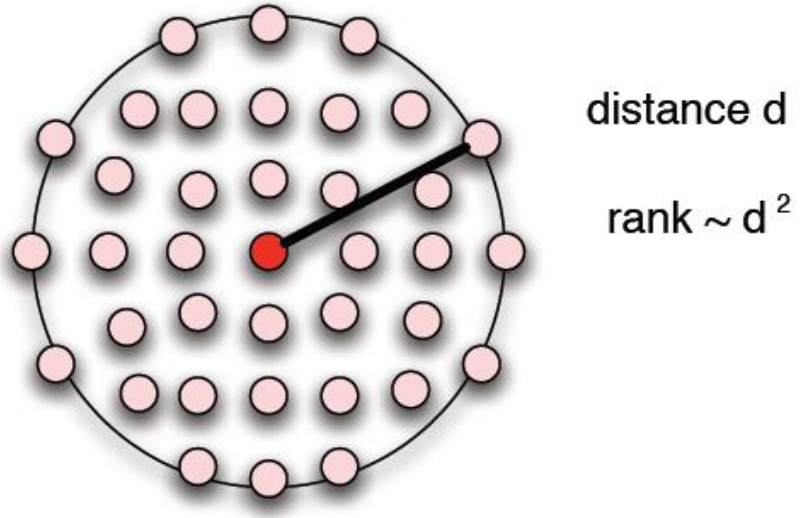


Linking by rank

- Link to the r-th closest neighbor with probability $P(x \to y) \sim r^{-1}$
  - In the case of uniform distribution, $P(x \to y) \sim d^{-2}$

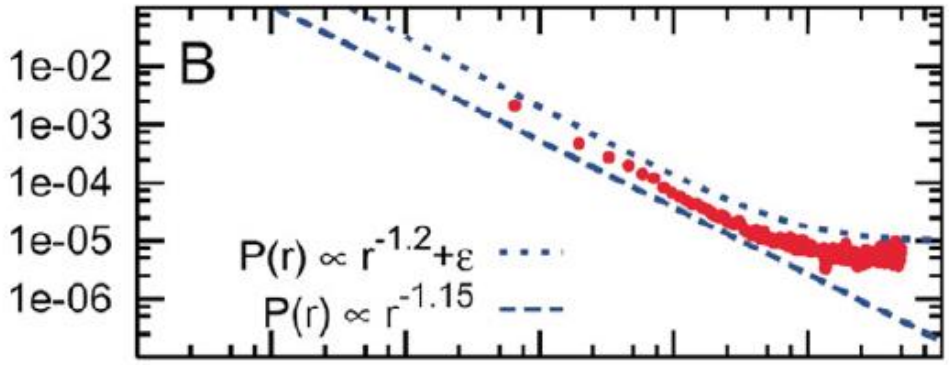


(a) *w is the 7th closest node to v.*

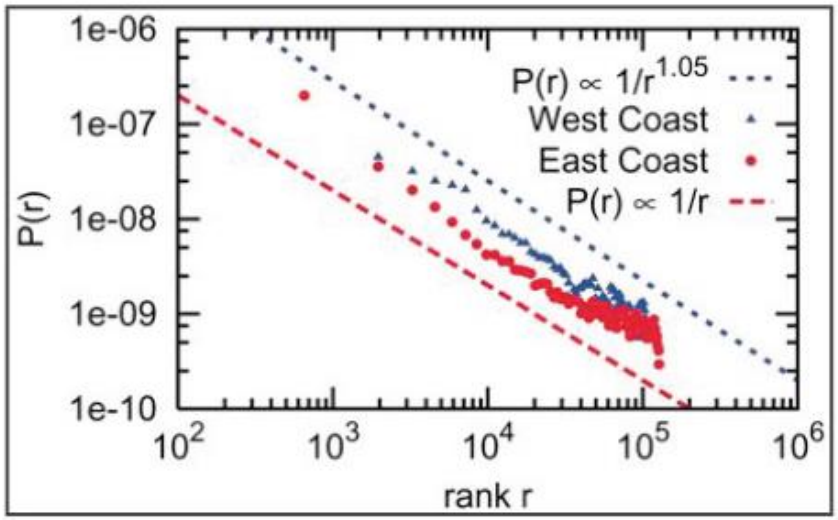


(b) *Rank-based friendship with uniform population density.*

Live Journal measurements



(a) *Rank-based friendship on LiveJournal*



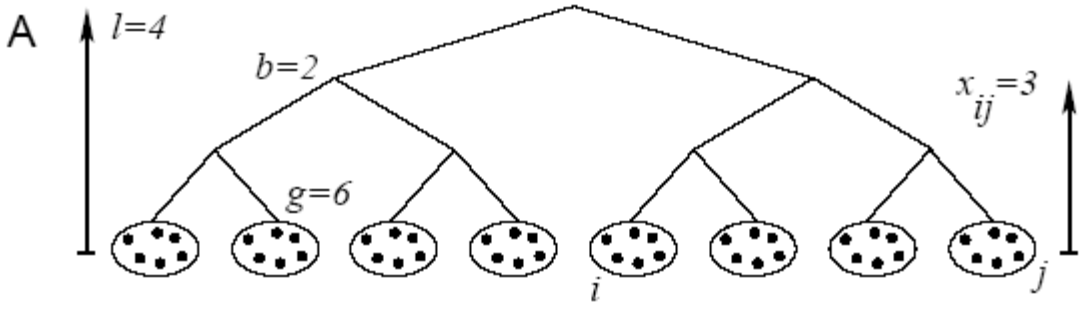
(b) *Rank-based friendship: East and West coasts*

Figure 20.10: The probability of a friendship as a function of geographic rank on the blogging site LiveJournal. (Image from [277].)

- Replicated for other networks as well (FB)
- Is there a mechanism that drives this behavior?

Other models

- Lattice captures geographic distance. How do we capture social distance (e.g. occupation)?
- Hierarchical organization of groups
  - distance h(i,j) = height of Least Common Ancestor

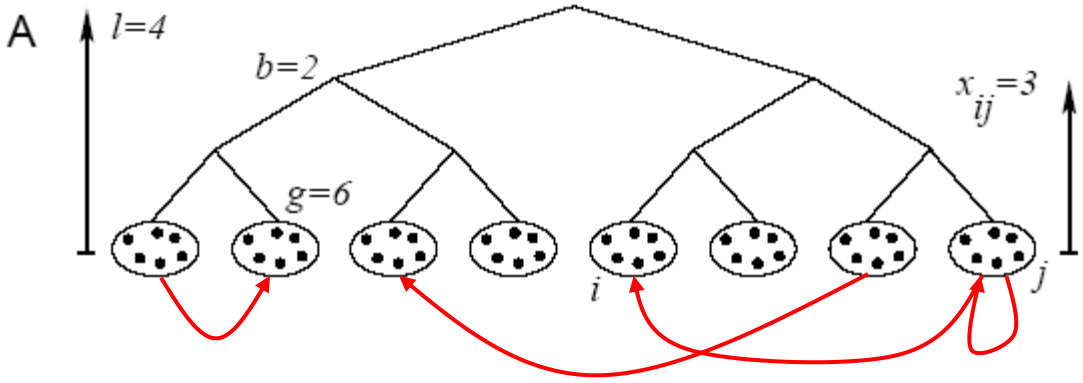


Other models

- Generate links between leaves with probability

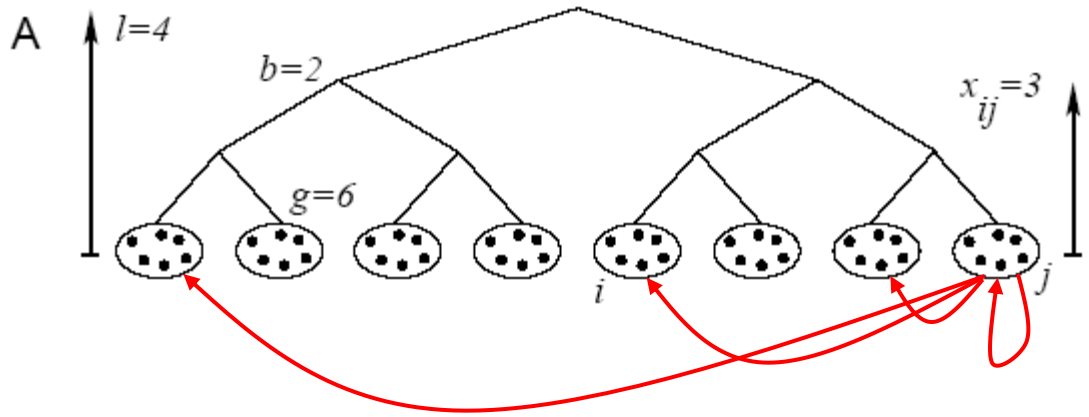
$$P(x \to y) \sim b^{-\alpha h(x,y)}$$

  - b=2 the branching factor



Other models

- Theorem: For $\alpha=1$ there is a polylogarithimic search algorithm. For $\alpha\neq1$ there is no decentralized algorithm with poly-log time
  - note that $\alpha=1$ and the exponential dependency results in uniform probability of linking to the subtrees



Generalization

- Social Distance: size of the smaller group that contains two users
- $P(x \to y) \sim d^{-1}$

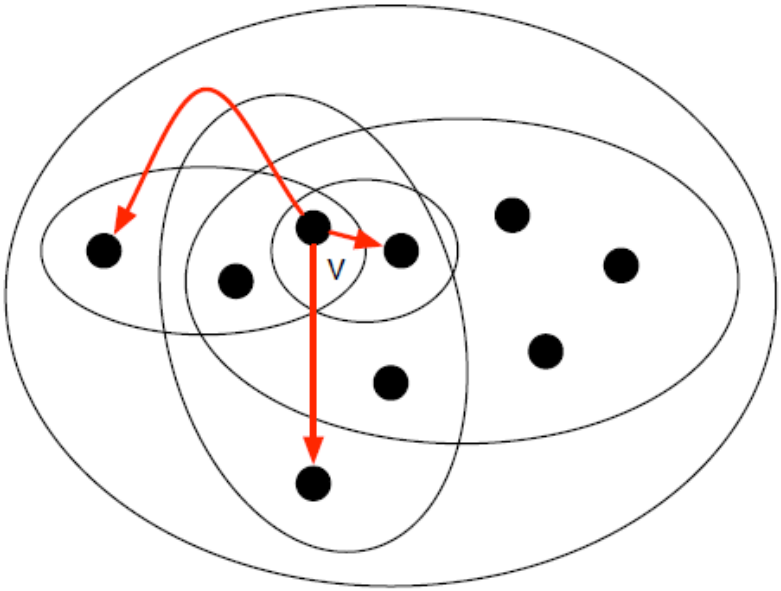


Figure 20.11: When nodes belong to multiple foci, we can define the social distance between two nodes to be the smallest focus that contains both of them. In the figure, the foci are represented by ovals; the node labeled v belongs to five foci of sizes $2, 3, 5, 7$, and 9 (with the largest focus containing all the nodes shown).

Doubling dimension

- A point set X has doubling dimension λ if any set of points in X that are covered by a ball of radius r can be covered by 2^{λ} balls of radius r/2.
- Practically, for any point x, if $N(x, r)$ is the number of points within distance r of x, then $N(x, 2r) = 2^{\lambda} N(x, r)$
  - According to what we have seen so far, to have logarithmic search time we need to add random links with probability $P(u, v) \approx d(u, v)^{-\lambda}$

Small worlds with nodes of different status

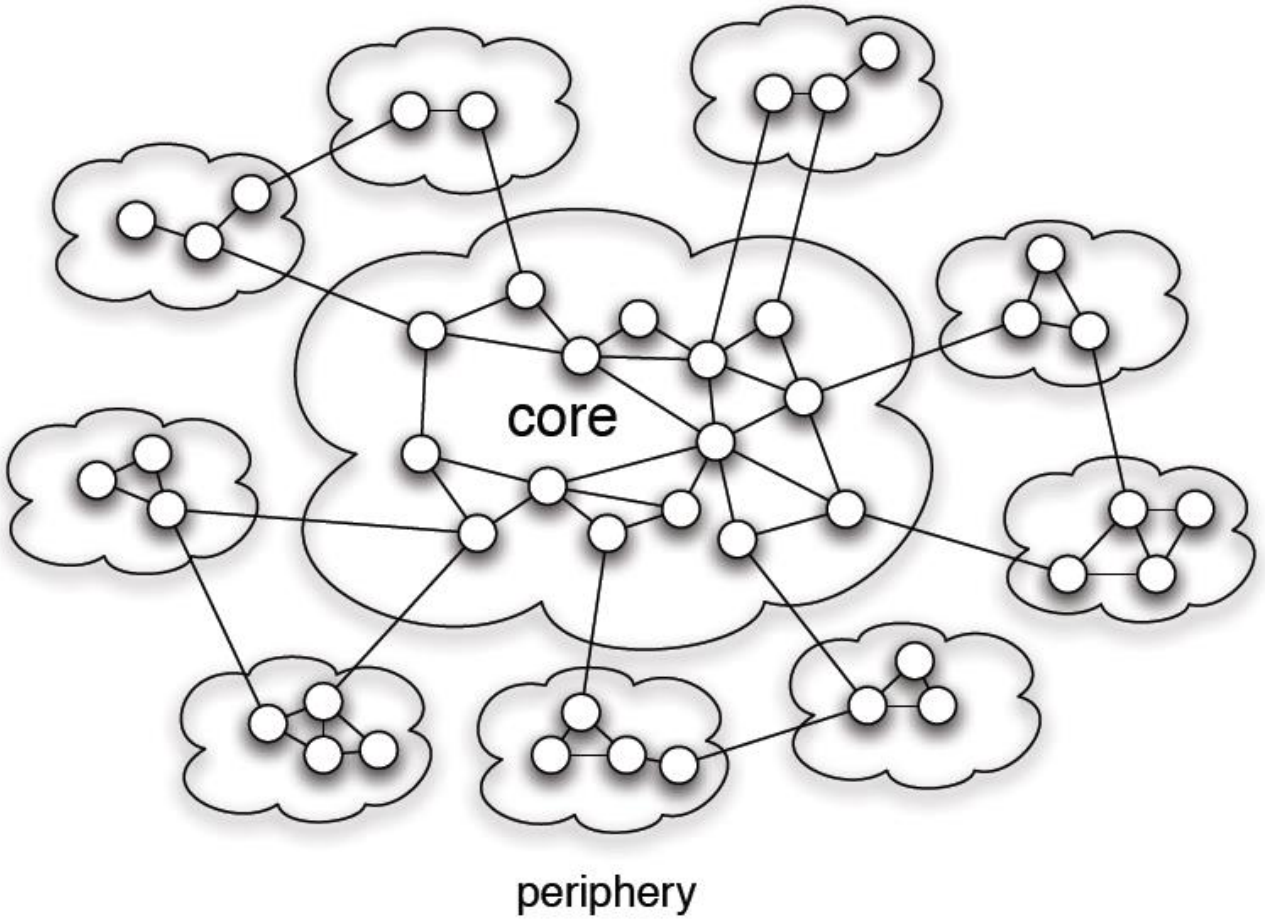
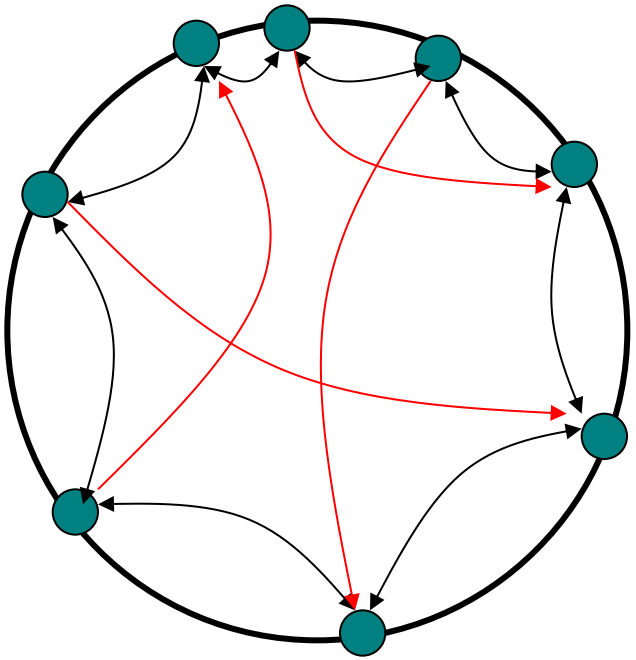


Figure 20.13: The core-periphery structure of social networks.

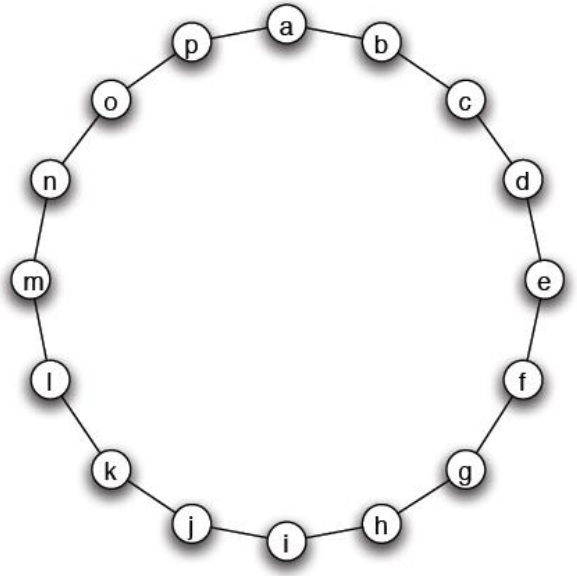
Application: P2P search -- Symphony

- Map the nodes and keys to the ring
  - Assign keys to the closest node
- Link every node with its successor and predecessor
- Add k random links with probability proportional to 1/(dlogn), where d is the distance on the ring
- Lookup time $O(\log^2 n)$
- If k = logn lookup time O(logn)
- Easy to insert and remove nodes (perform periodical refreshes for the links)

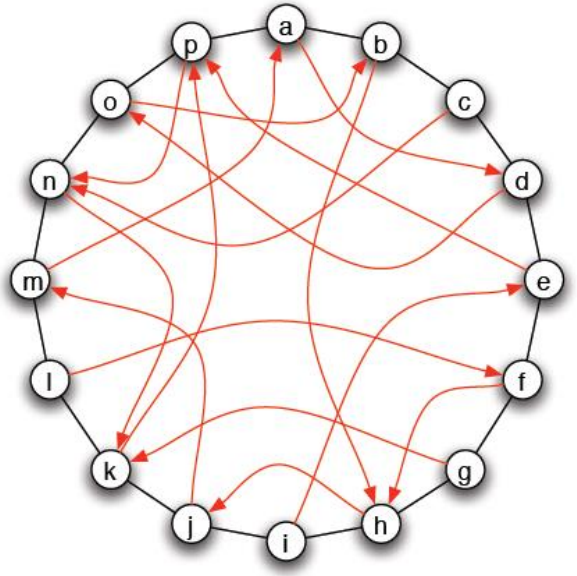


Proof of Kleinberg's theorem

- We will consider the ring and clustering exponent $q = 1$



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



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

Proof of Kleinberg's theorem

- Game plan:
  - Break the path from s to t into phases:
    - In phase j we are at distance $(2^j, 2^{j+1}]$ from t
  - When transitioning between phases we cut the remaining distance from s to t in half
    - There are $\log n$ phases
  - Show that the expected time spent in each phase is $O(\log n)$
  - Total time: $O(\log^2 n)$

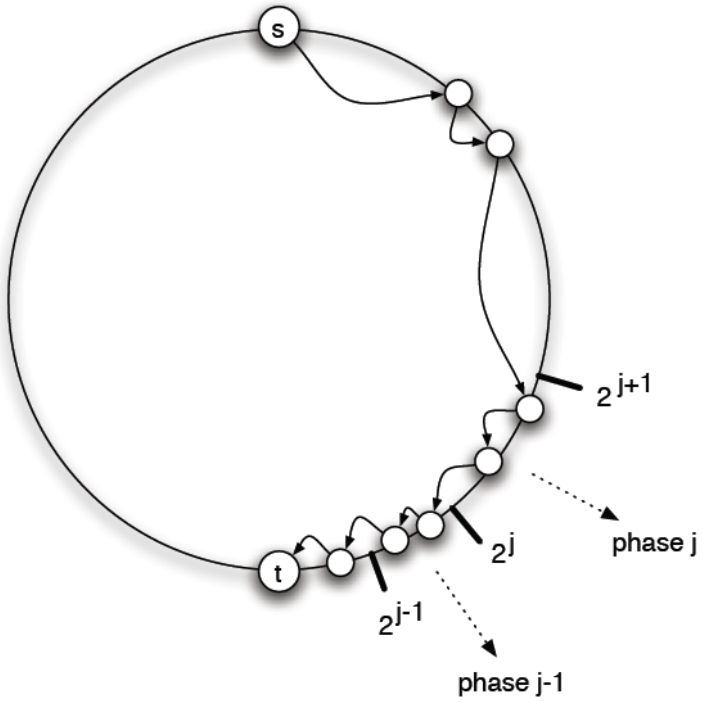


Figure 20.16: We analyze the progress of myopic search in *phases*. Phase j consists of the portion of the search in which the message's distance from the target is between 2^j and 2^{j+1}.

Normalization constant

$$Z = \sum_y d(x,y)^{-1} = 2\left(1 + \frac{1}{2} + \cdots + \frac{1}{n/2}\right)$$
$$\leq 2\left(1 + \ln\frac{n}{2}\right)$$
$$\leq 2\,(1 + \log n/2) = 2\log n$$

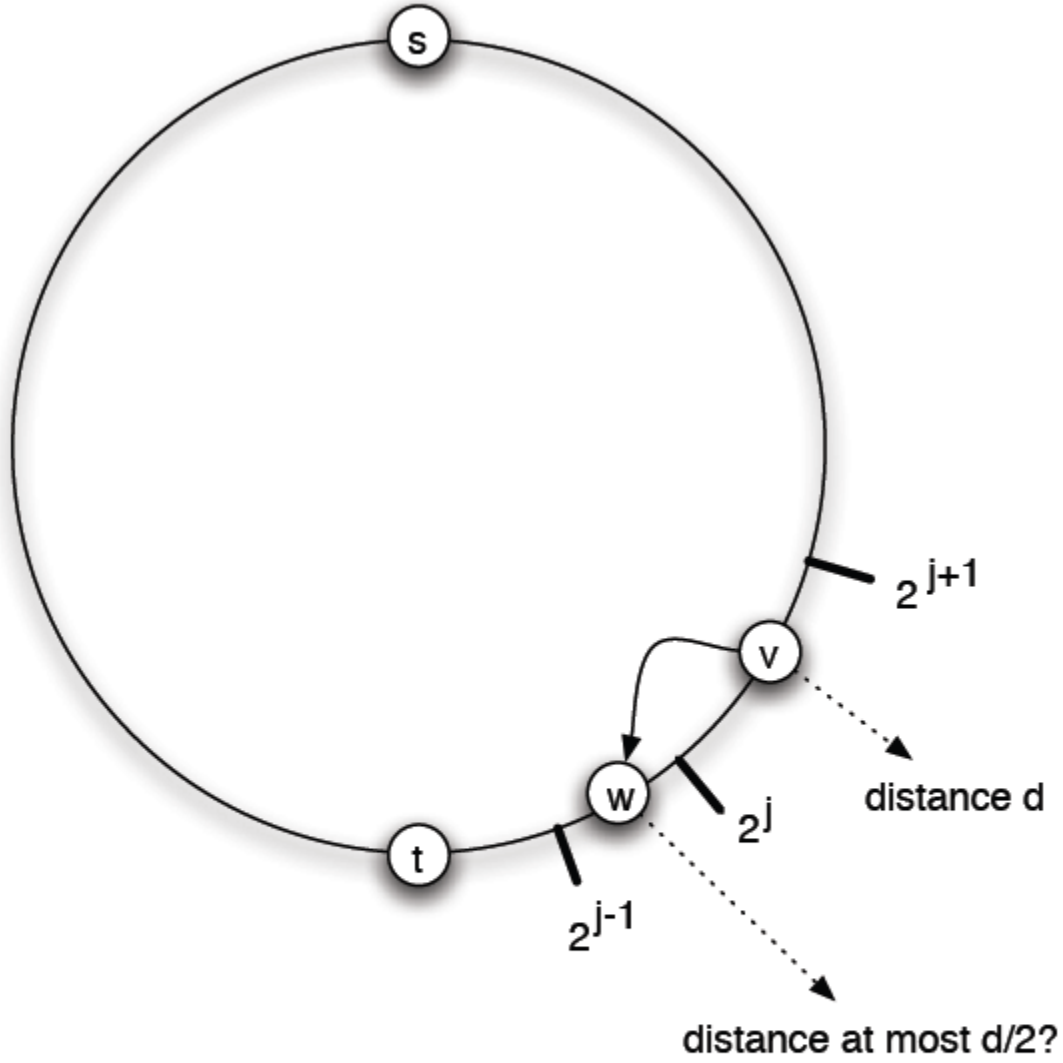


Figure 20.18: At any given point in time, the search is in some phase j, with the message residing at a node v at distance d from the target. The phase will come to an end if v's long-range contact lies at distance $\leq d/2$ from the target t, and so arguing that the probability of this event is large provides a way to show that the phase will not last too long.

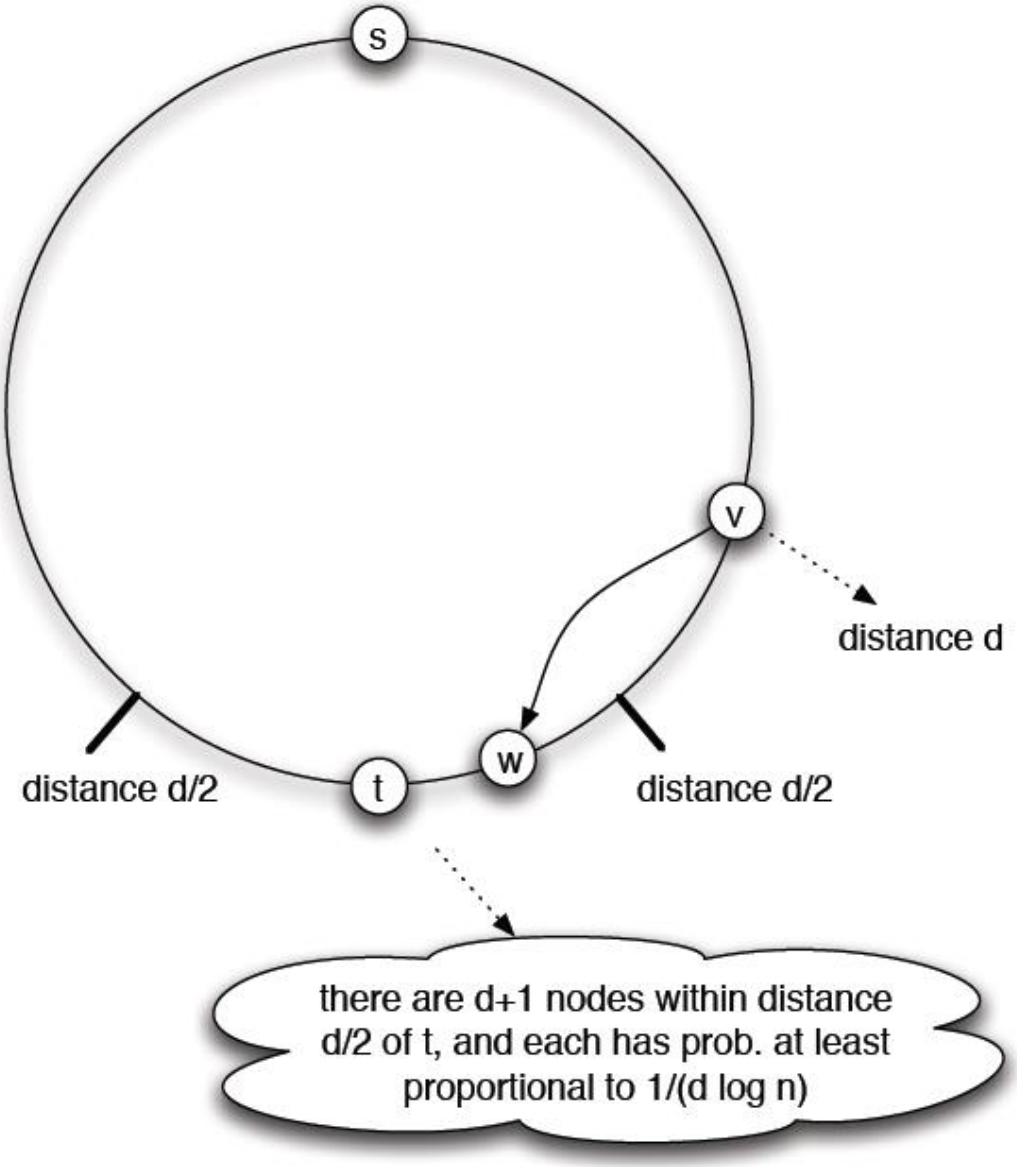


Figure 20.19: Showing that, with reasonable probability, v's long-range contact lies within half the distance to the target.

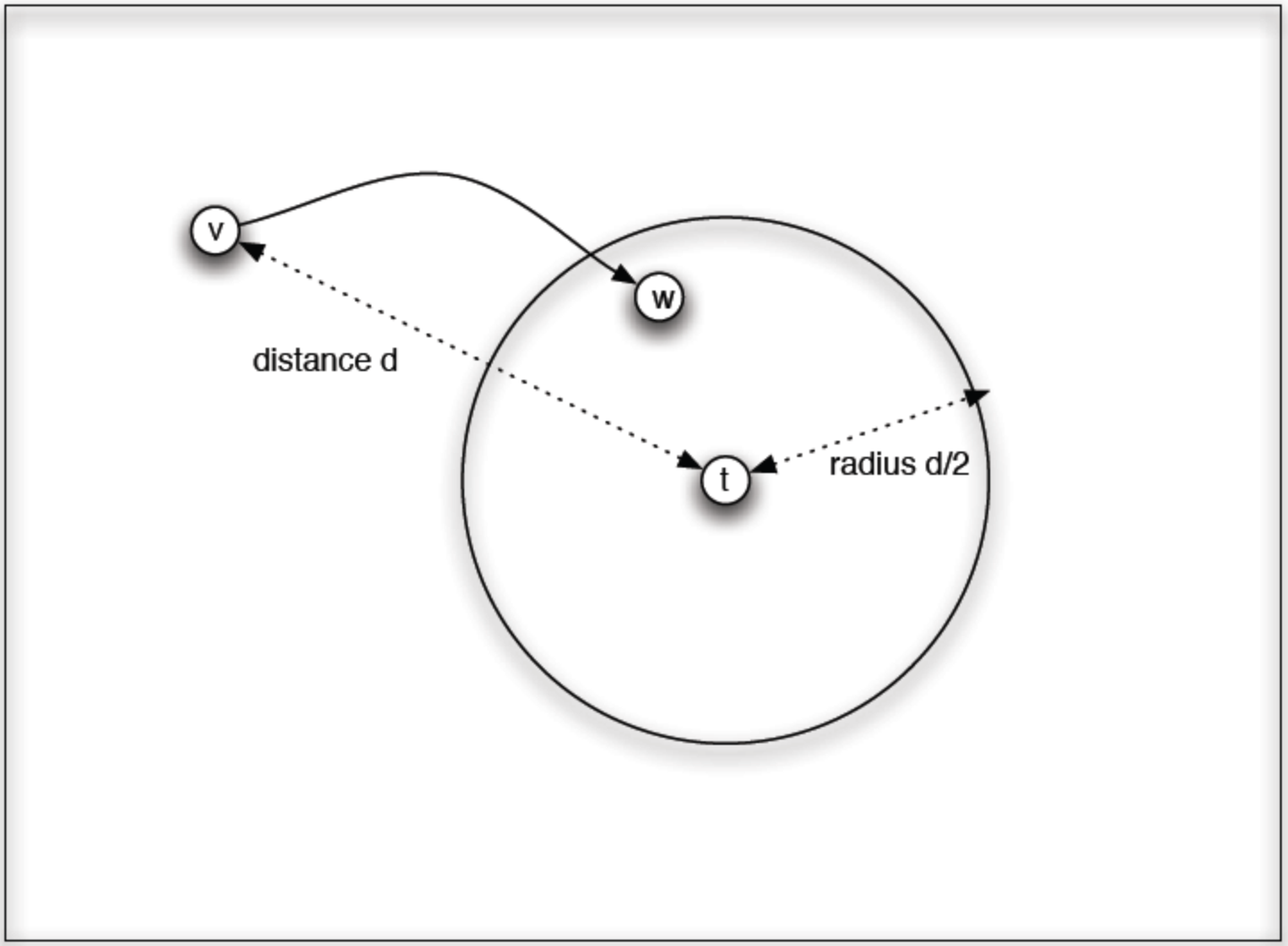


Figure 20.20: The analysis for the one-dimensional ring can be carried over almost directly to the two-dimensional grid. In two dimensions, with the message at a current distance d from the target t, we again look at the set of nodes within distance $d/2$ of t, and argue that the probability of entering this set in a single step is reasonably large.

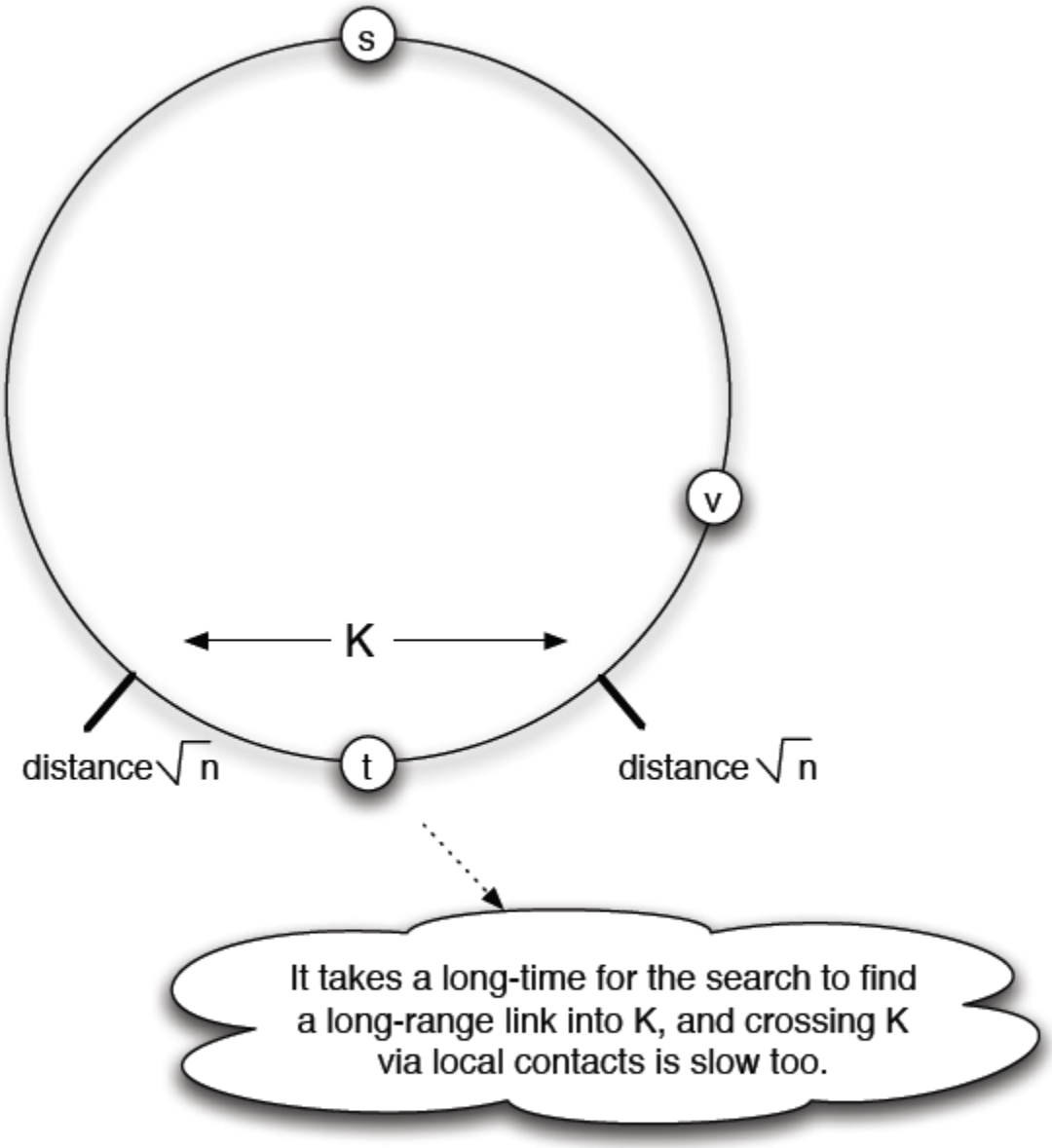


Figure 20.21: To show that decentralized search strategies require large amounts of time with exponent $q = 0$, we argue that it is difficult for the search to cross the set of $\sqrt{n}$ nodes closest to the target. Similar arguments hold for other exponents $q < 1$.