Social Network:
The Small-World Phenomenon and Decentralized Search

Advanced Computer Networks

Summer Semester 2012

Recap: Information Cascading Model

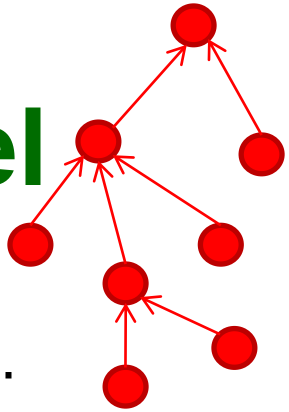
- Consider an urn with 3 marbles. It can be either:
  - Majority-blue: 2 blue, 1 red
  - Majority-red: 1 blue, 2 red
- Each person wants to best guess whether the urn is majority-blue or majority-red
- Experiment: making decision one by one
- Analysis: Bayes' Rule (posterior probability)

$$\Pr[A \mid B] = \frac{\Pr[A] \cdot \Pr[B \mid A]}{\Pr[B]}.$$

- Decision: based on the best probability

$$Pr[majority - blue|blue, blue, red] = \frac{2}{3} \geq \frac{1}{2}$$

Recap: Rich Get Richer Model



- Creation of links among Web pages
  - Pages are created in order, and named 1; 2; 3; …;N.
  - When page j is created, it produces a link to an earlier Web page i according to:
  - 1) With prob. p (0<p<1), j links to i chosen uniformly at random (from among all earlier nodes)
  - 2) With prob. 1-p, node j links to node u with prob. proportional to the degree of u
- Major results: let q=1-p, for degree k, by estimation

$$Pr\{x \geq k\} = [\frac{q}{p} \cdot k + 1]^{-1/q}$$

$$F\{x\} = Pr\{x < k\} = 1 - Pr\{x \geq k\}$$

$$Pr\{x = k\} = F'(x) = \frac{1}{p}[\frac{q}{p} \cdot k + 1]^{-(1+1/q)}$$

Power-Law!

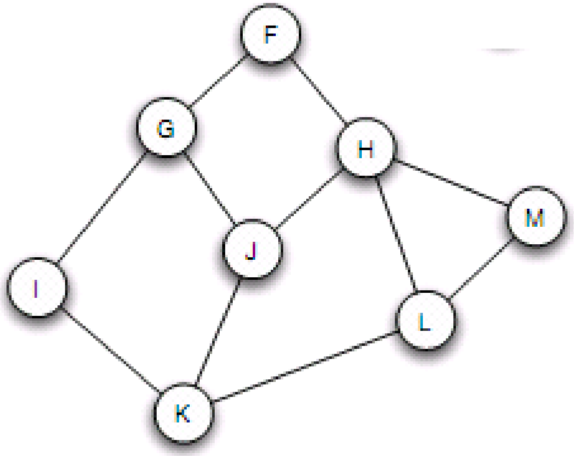
Key Network Properties

How to Characterize Networks?

- How many neighbors does a node have?
  - Degree distribution
  - Power-law for many social networks
- How far apart are nodes in the network?
  - Distance (the shortest path)
  - Network diameter
  - Average path length
- How close a set of nodes connect with each other?
  - Community
  - Clustering coefficient

Path Length

- **Distance**: the number of edges along the shortest path connecting the nodes
  - If two nodes are disconnected, the distance is infinite
- **Diameter**: the maximum distance between any pair of nodes in the graph
- **Average path length**:



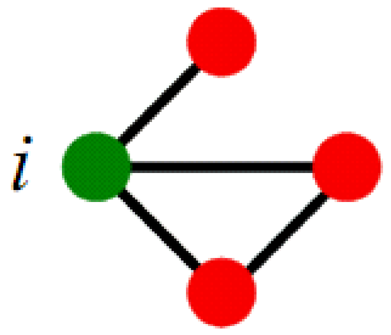
Clustering Coefficient

- Evaluate how the neighbors of a node are connected
- For node i with degree k, assume the number of edges between the neighbors of i is e, the clustering coefficient of i is

$$C_i = \frac{e}{k(k-1)/2}$$

- Average clustering coefficient

$$C = \frac{1}{N}\sum_{i}^{N} C_i$$



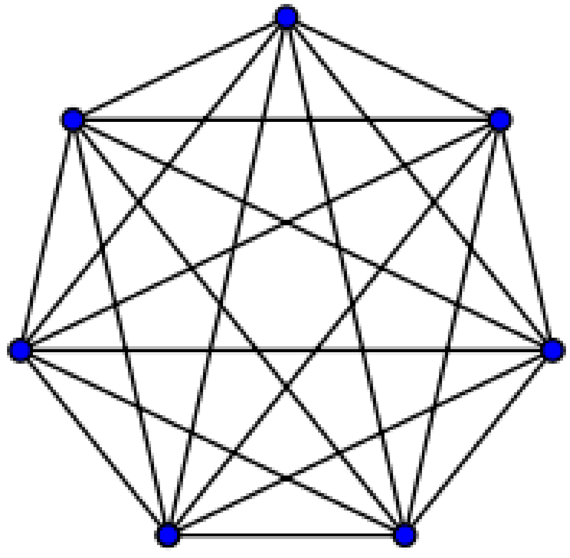
$C_i = 1/3$

Key Network Properties

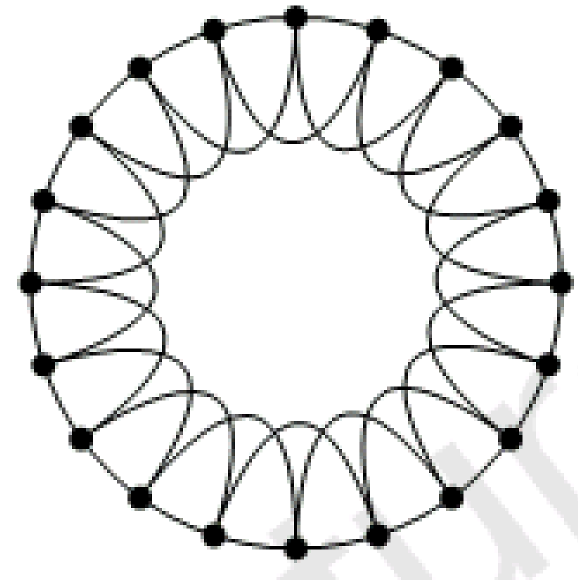
- Degree distribution: $P(k)$
- Path length: h
- Clustering coefficient: C

Complete Graph

- Degree distribution: P(k)=N-1
- Path Length:
  - Diameter: 1
  - Average path length: 1
- Clustering coefficient
  - C=1
  - Average clustering coefficient: 1



Regular Lattice



- Degree distribution:

$$P(k) = \begin{cases} 1, & k = 4 \\ 0, & \text{otherwise} \end{cases}$$

- Path length:
  - Diameter: $h_{max} = \frac{N}{4}$
  - Average: for node, its distance to other nodes are: 1, 1, 2, 2, 3, 3, …, N/4, N/4.
  - So

$$h_{avg} = \frac{2 \times (1 + 2 + \cdots + N/4)}{N/2} = \frac{2\frac{(1+N/4)*N/4}{2}}{N/2} = 1/2 + \frac{N}{8}$$

- Clustering coefficient $C_i = \frac{e}{k(k-1)/2}$
  - C=2*3/(4*3)=1/2 for N>6
- Summary: constant degree, constant clustering coefficient, but average path is O(N)

Random Graph

- Degree distribution: Binomial distribution

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

- Average path length: $O(\log n)$

- Clustering coefficient: $C = p = \bar{k}/n$

The Small-World Phenomenon

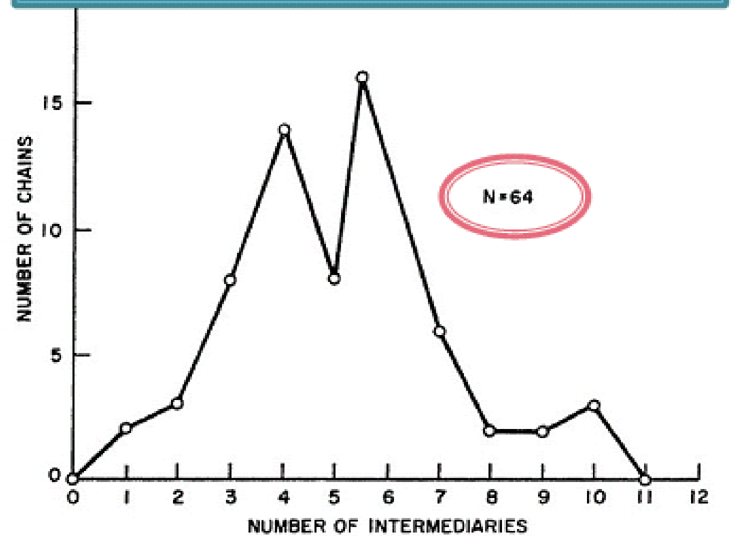
Six Degrees of Separation

- What is the typical shortest length between any two people in human society?
  - Global measurement is impossible
  - Sampling
- Experiment
  - Milgram 1967
  - Idea: ask randomly chosen “starter” individuals to try to forward a letter to a designated “target” person

Milgram's Experiment [1967]

- Procedure
  - The target person
    - A stockbroker who worked in Boston and lived in Sharon, Massachusetts
  - The starting person
    - Randomly picked 300 people in Omaha, Nebraska and Wichita, Kansas
  - The target's name, address, occupation, and some personal information are provided
  - Rules: "If you do not know the target person on a personal basis, do not try to contact him directly. Instead, mail this folder … to a personal acquaintance who is more likely than you to know the target person … it must be someone you know on a first-name basis".
  - The names of the person who forward the letter are attached

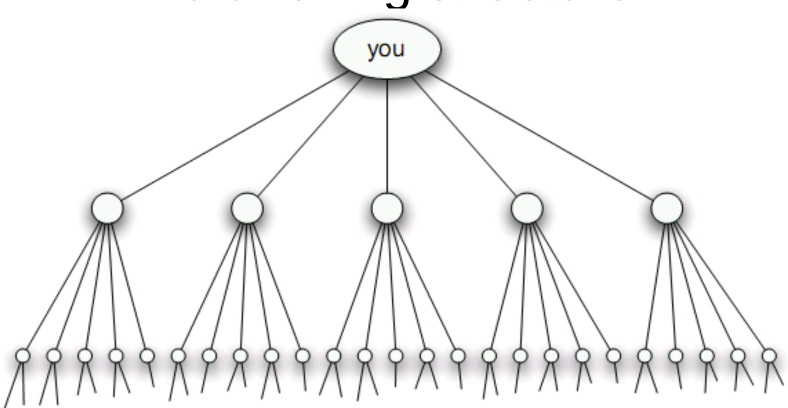
Milgram's small world experiment

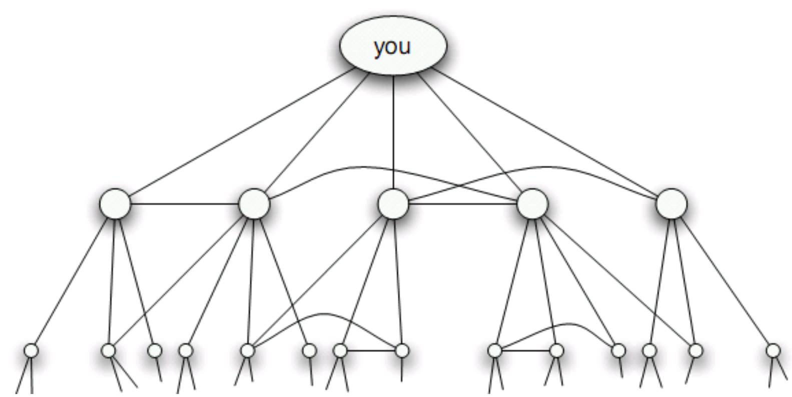


- How many steps did it take?
  - 64 letters reached the target
  - It took 6.2 steps on average
- Short paths exist! -- Six Degrees of Separation
- Similar results are verified in other social networks like actor network, email network, who-talks-to-whom network (MSN), Facebook …
- Two facts
  - Short paths are there in abundance
  - People without global "map" of the network are effective at collectively finding these short paths (How to do decentralized search?)

A Simple Explanation

- Suppose each person knows 100 other people on a first-name basis
  - Step 1: reach 100 people
  - Step 2: reach 100*100 people
  - …
  - Step 5: reach 100^5= 10 billion people
  - Ref: the world population is 7.019 billion (Wiki, 2012)
- The numbers are growing by powers of 100
- But it is not true for real network!!!
  - Triadic relationships are common
  - Social network is highly clustered, not the kind of massively branching structure.



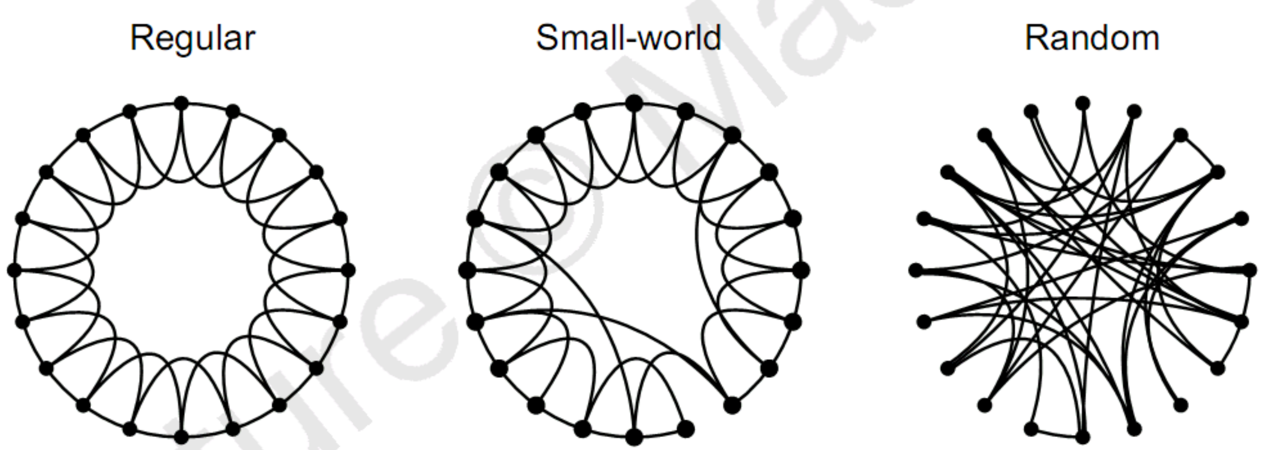


Regular Network vs Small-World Network vs Random Network

- Regular network: high clustering, high diameter
- Random network: low clustering, low diameter
- Question
  - Is there a network inbetween the regular network and random network, with high clustering coefficient and low average path length?
- Small-World network: high clustering, low diameter

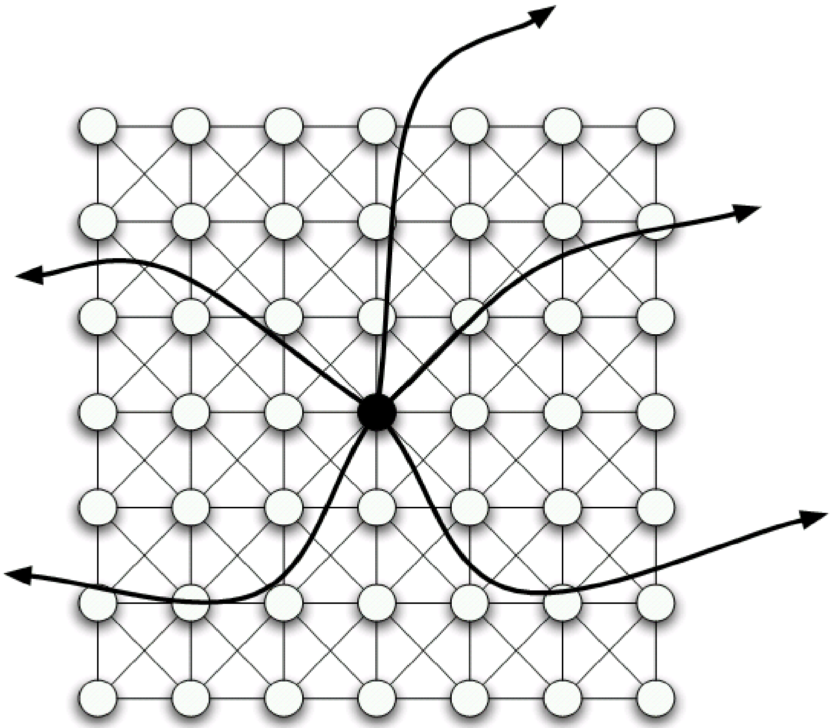
The Small-World Model

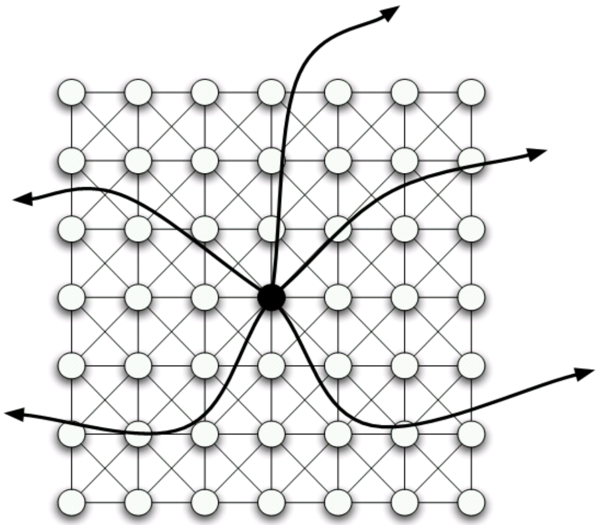
- Can we make up a simple model that exhibits both of the features: many closed triads, but also very short paths?
- **One-deminsional Model (Watts-Strogatz)**
- Starting from a ring lattice with n vertices and k edges per vertex.
  - Regular network with high clustering coefficient
- We rewire each edge at random with probability p (0≤p≤1).
  - p=0: regular network
  - p=1: random network
  - Randomizing the network, lowering average path length



The Watts-Strogatz Model

- **The two-dimensional model: grid**
- Two kind of links
  - Regular links: Links to the other nodes within a radius of up to r grid steps
  - Random: Links to k other remote nodes

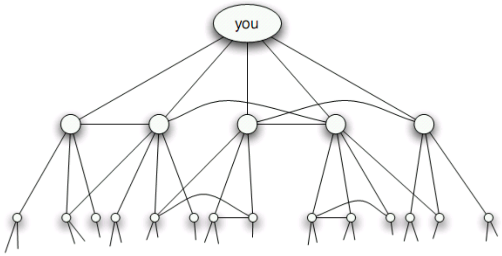




- High clustering

$$C_i \geq 2*12/(8*7) \geq 0.43$$

- Low diameter: short path exists with high probability
  - Since the k remote nodes are random and they barely know each other
    - For each step, at lease k new nodes are reached
    - The numbers are growing by powers of k
  - Still, short path achieves, the diameter is O(logn)



Extension

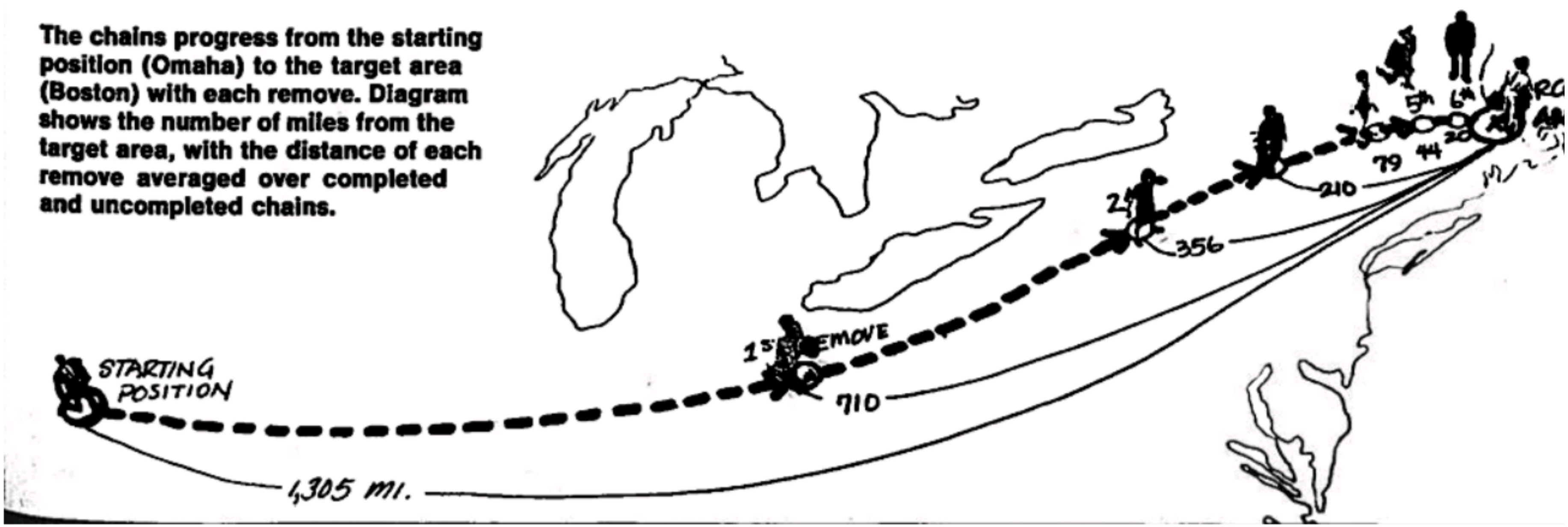
- Short path still exists even for very small amount of randomness
- For example, instead of allowing each node to have k random friends, we only allow one out of every k nodes to have one random friend
  - We can conceptually group k*k subsquares of the grid into “towns”
  - It will be similar: each town links to k other towns
  - Short path in towns -> short path in people

Small World: Summary

- A network between regular network and random network
- It has high clustering and low diameter
  - Clustering efficient: much larger than random network
  - Diameter: almost equal to random network
- The Watts Strogatz Model
  - Introducing a tiny amount of random links is enough to make the world small, with short paths between every pair of nodes.

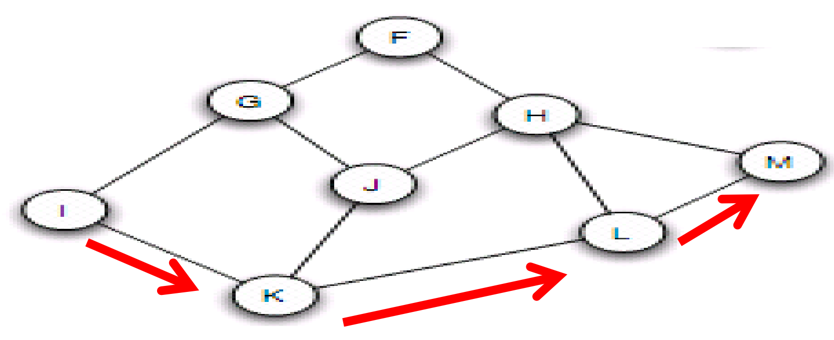
Decentralized Search

- Question
  - In a Small World network, how to find the short path between a pair of nodes?
    - Centralized strategy?
    - Flooding?
    - Milgram experiment: people collectively find short paths to the designated target -> decentralized search is possible

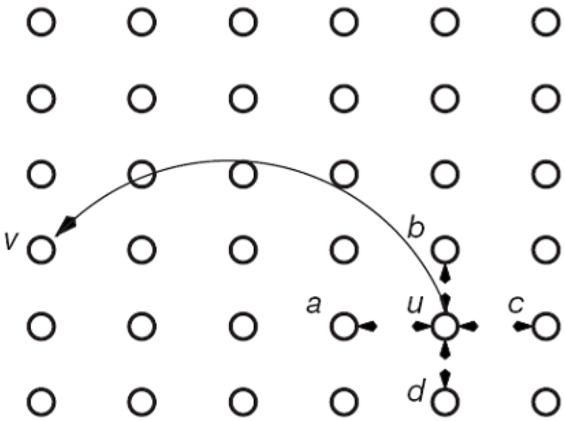


Decentralized Search

- Node s sending a message to destination t
  - s only knows locations of its friends and locations of the target t
  - s only has local information, it does not know links of other nodes
- Principle: s send the message to its friend who is the closest to t
- Search path length: the number of steps to reach t



A General Network Model



- One dimension: A ring
- Two dimension: A grid
- Each node has only one long link
- The probability of a long link from u to v is:

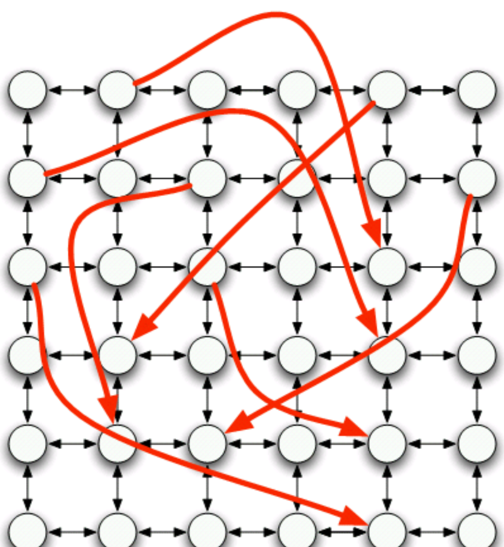
$$Pr\{u \to v\} \sim d(u,v)^{-q}$$

- Where d(u,v) is the distance (grid steps) between node u and v, and q is a parameter

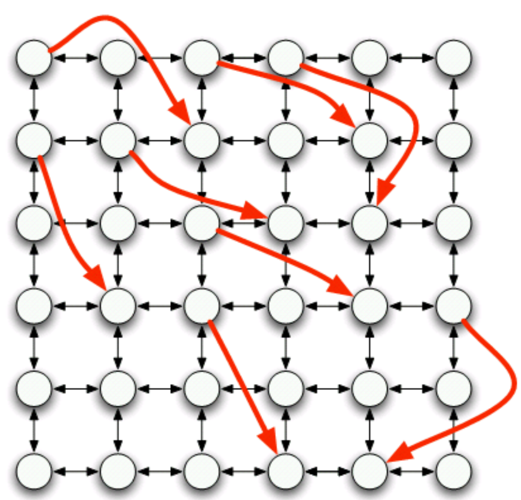
Choosing the parameter q

$$Pr\{u \to v\} \sim d(u,v)^{-q}$$

- Different q yields different networks, which have different shortest path lengths
- q=0: equals to the Watts-Strogatz model
- q-> +∞: only links to nearby nodes



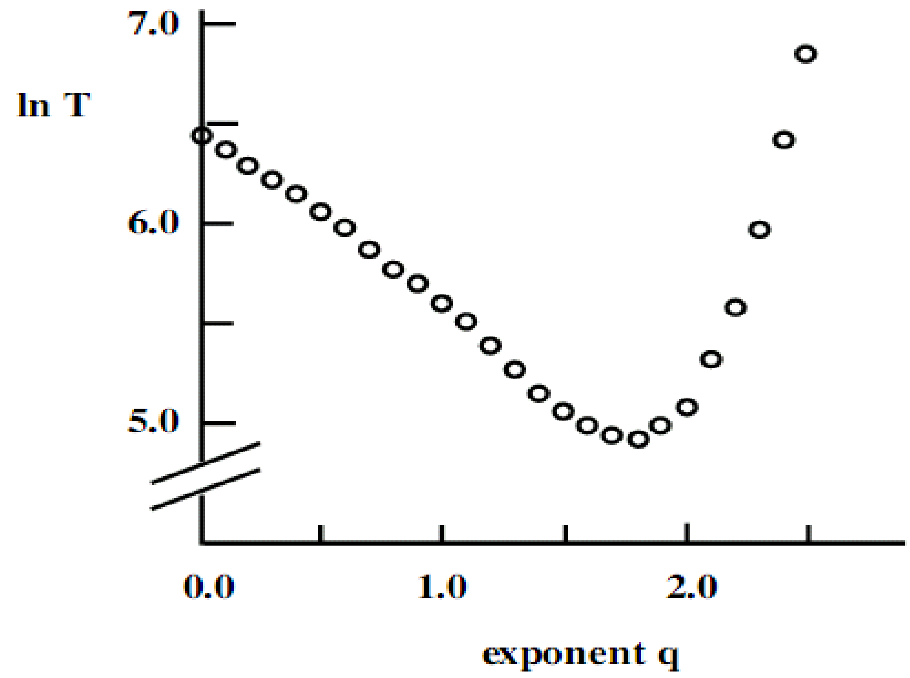
Too random!



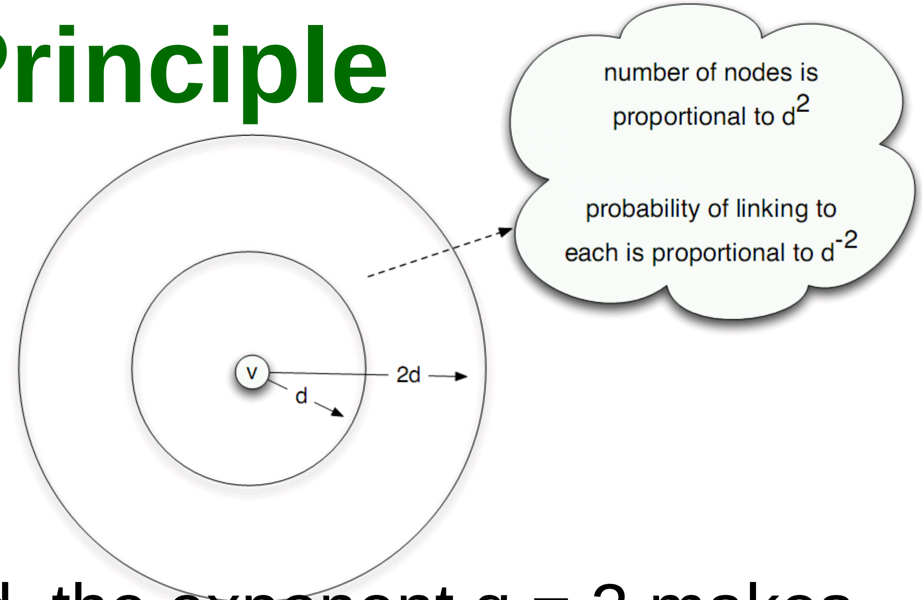
Not random enough!

What is the best value of q?

- Is there a value of q, making the search path achieves the shortest?
- Experiment on a two-dimensional grid
  - q≈2



Inverse-Square Principle



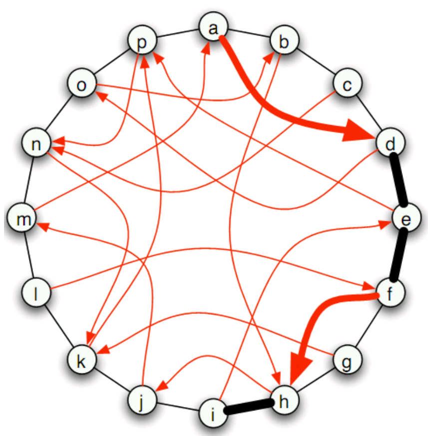
- For a two-dimensional grid, the exponent q = 2 makes it best for decentralized search

$$Pr\{u \to v\} \sim d(u,v)^{-2}$$

  - Guess: for d-dimensional, q=d!
- Rough explanation
  - The total number of nodes in an area is proportional to d^2
  - The probability for v linking to the nodes is proportional to d^{-2}
  - They cancel out -> making the probability from v to any other node in the area is independent of d

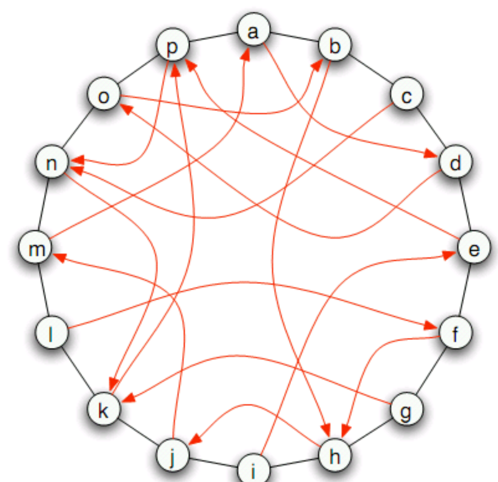
Analysis the Model in 1-dimension

- Nodes are arranged in a ring.
- For 1 dimension, p=1 is the best $Pr\{u \rightarrow v\} \sim d(u,v)^{-1}$
- Each node knows only local information, performing decentralized search
- Search strategy: Myopic search
  - When a node v is holding the message, it passes it to the contact that lies as close to t on the ring as possible
  - Not guarantee to be shortest path
- Question: what is the expected length of search path?

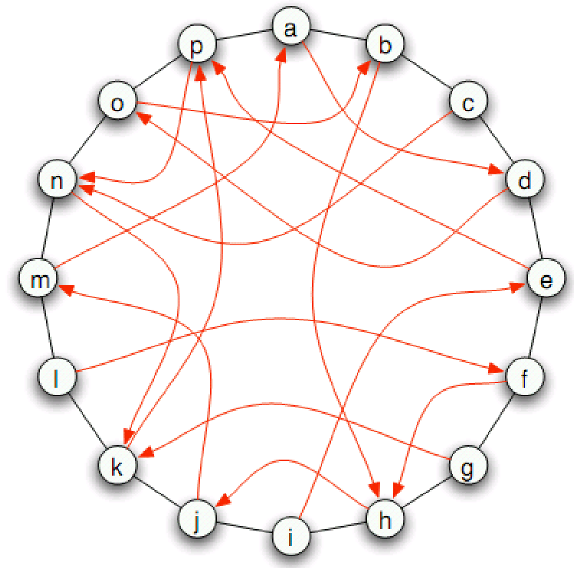


- Claim: for q=1 in 1-dimensional model, we can get from s to t in O(log(n)^2) steps.
- Proof:
- Normalization: $Pr\{u \to v\} \sim d(u,v)^{-1}$
  - Let $$Z = \sum_{i \neq u} d(u,i)^{-1}$$
  - The probability of linking from node u to v is:

$$Pr\{u \to v\} = \frac{d(u,v)^{-1}}{Z}$$



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



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

- Since

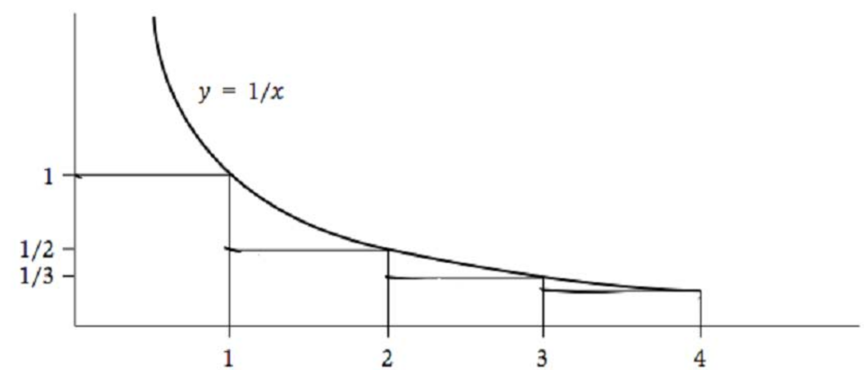
$$Z = \sum_{i \neq u} d(u,i)^{-1}$$

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

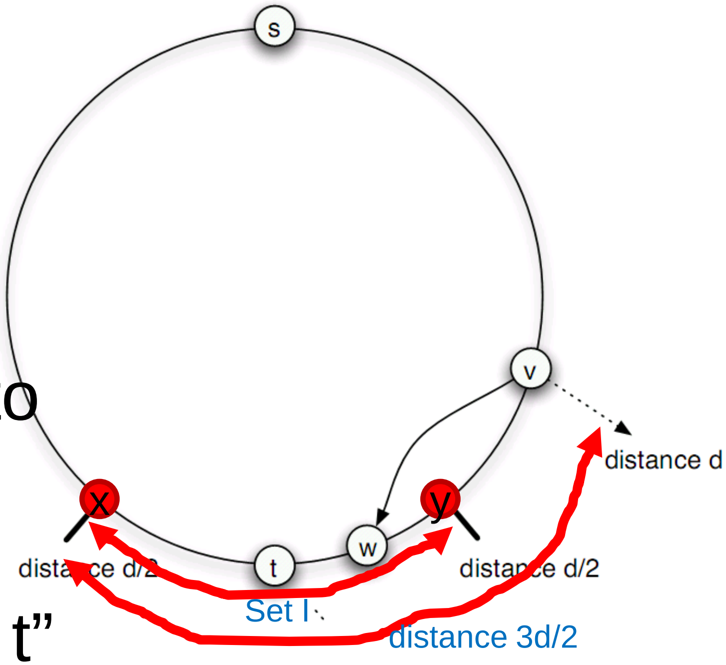
$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{k} \leq 1 + \int_1^k \frac{1}{x} dx = 1 + \ln k.$$

- We have

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



computer NETWORKS

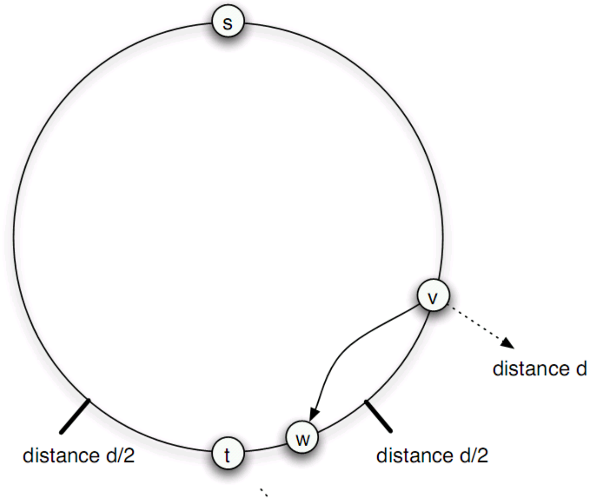


- For a node v, assume its distance to destination t is d, when will the message enters d/2 of t?
- Let I = “the set of nodes with d/2 of t”
  - The number of nodes in I is d+1

$$Pr\{v\ points\ to\ I\} = \sum_{j \in I} Pr\{v \to j\} = \sum_{j \in I} \frac{d(v,j)^{-1}}{Z}$$

$$= \frac{1}{Z} \sum_{j \in I} \frac{1}{d(v,j)} \geq \frac{1}{Z}(d+1)\frac{1}{d(v,x)} \geq \frac{1}{Z} d \frac{2}{3d} \geq \frac{2}{3Z}$$

- Since $Z \leq 2ln(n)$
- We have $$Pr\{v\ points\ to\ I\} \geq \frac{1}{3ln(n)}$$



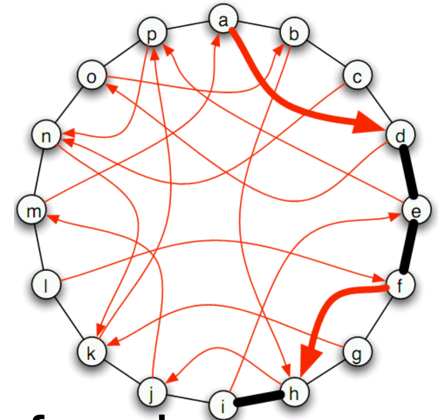
- We have

$$Pr\{v \ points \ to \ I\} \geq \frac{1}{3ln(n)} = O(\frac{1}{ln(n)})$$

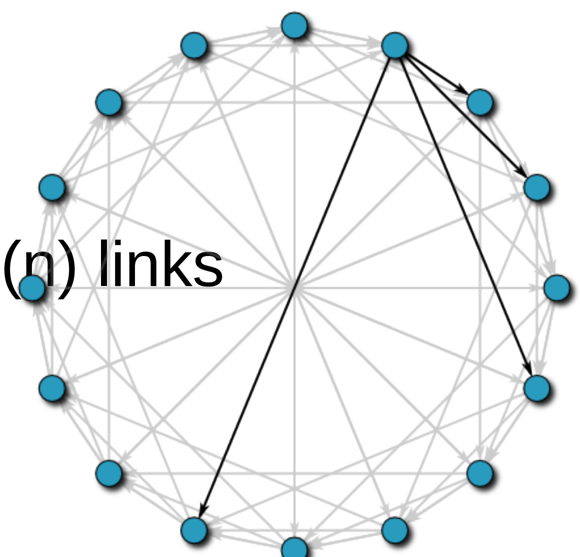
- It means within O(ln(n)) steps, we can get into I from v (the distance is halved!)
- Distance can be halved at most $\log_2(n)$ times, so the expected time from s to t is

$$\mathrm{O}(\ln(n)\cdot\log_2(n)) = \mathbf{O}(\log(n)^2)$$

Summary



- In 1-dimenstional ring structure
  - Each node knows only local information, performing decentralized search
  - Search strategy: Myopic search
  - p=1 achieves the shortest search path length
  - Expected search path: $O(\log(n)^2)$
- Compare with P2P searching?
  - Chord
  - Each node has a FingerTable with log(n) links
  - The search path length is O(log(n)).



Analysis in Two Dimensions

- For 2-dimensional grid, q=2 achieves the best for decentralized searching
  - For n-dimensional, should be q=n.
- Analysis is similar to 1-demensional case
  - Normalization: z is still O(ln(n)) $Z = \sum_{i \neq u} d(u,i)^{-2}$
  - The number of nodes within d/2 of the target is $O(d^2)$
  - The probability v link to one node in I is $O(1/d^2Z)$
  - The probability of halving the distance is:

    $O(d^2) * O(1/d^2Z) = O(1/Z)$ (d is canceled out!)
    - Similar, for n-dimensional, letting q=n will cancel out d
  - The expected steps to halve the distance is O(Z)=O(ln(n))
  - The total expected steps from s to t is: $\log_2 n * O(\ln(n)) = O(\log(n)^2)$
- This is called Inverse-Square Principle $Pr\{u \rightarrow v\} \sim d(u,v)^{-2}$