Social and Information Networks

CSCC46H, Fall 2022

Lecture 5

Prof. Ashton Anderson
ashton@cs.toronto.edu

Logistics

A2 out tomorrow, due in 3 calendar weeks (including reading week)

Today

Six degrees of separation
Search in networks

Q: How connected are we?

Fundamental question in network science:
How **connected** are we?

We've already seen one answer: the vast majority of the world is at least connected somehow (giant component)

How socially "far apart" are people?

Q: How connected are we?

Vast majority of people are in the giant component

For any pair of people in the giant component, there is a friendship path between them

How long are these paths?

How long are real-world paths?

Think of a random person in the world:
Baker from India? Telemarketer from Tuvalu? Brazilian photographer? Farmer from Germany?

How many links in the chain from you to them? (network distance)

How long is the typical shortest path?

Milgram 1967 was the first to study this question
But back then, no explicitly recorded social network!

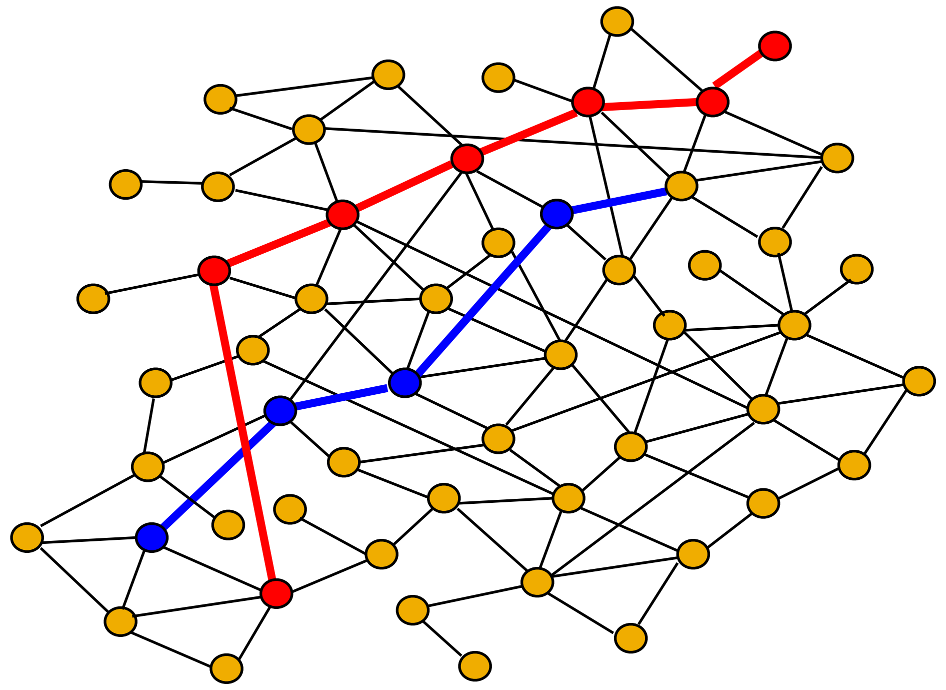
What would you do?

How long is the typical shortest path?

Milgram devised a clever experiment

–Picked ~300 people in Omaha, Nebraska and Wichita, Kansas
–Asked each person to try get a letter to a particular person in Boston (a stockbroker), but they could only send it to someone they know on a first-name basis
–The friends then send to their friends, etc.

How many steps did it take?

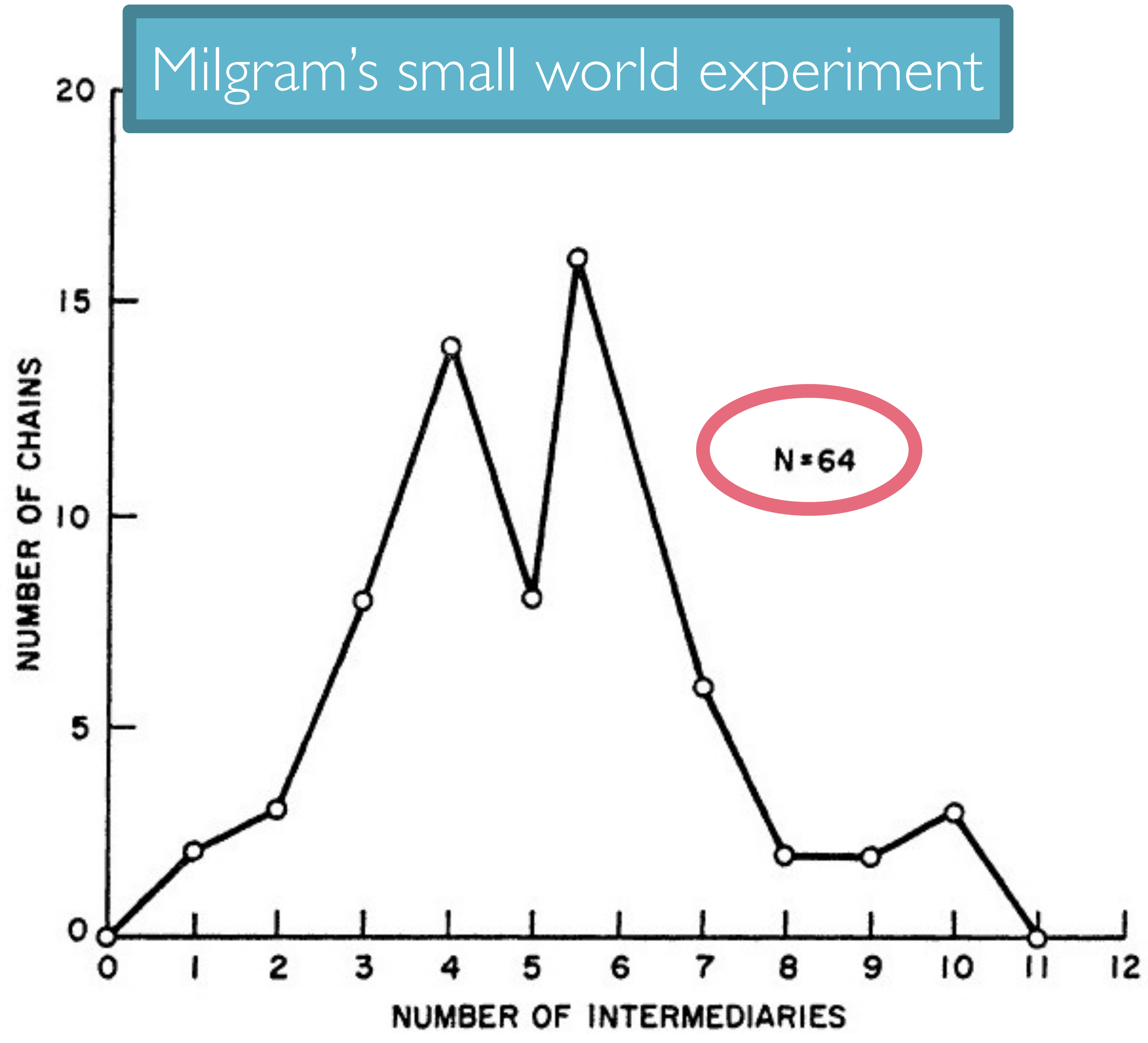


The Small-World Experiment

64 chains completed:

(i.e., 64 letters reached the target)

It took 6.2 steps on the average, thus **"6 degrees of separation"**



Further observations:

People who owned stock had shorter paths to the stockbroker than random people: 5.4 vs. 6.7

People from the Boston area have even shorter paths: 4.4


facebook
December 2010

Network Analysis Methodology: Six degrees of freedom

What are the basic properties properties of real social networks?

How can we model them?

Today:

- Milgram's experiment   <— **You are here**
- Measuring path lengths in real-world networks
- Comparing with a baseline: G_{np} model
- More realistic models: Watts&Strogatz model
- More realistic models: Kleinberg's Decentralized search

Six Degrees of Kevin Bacon

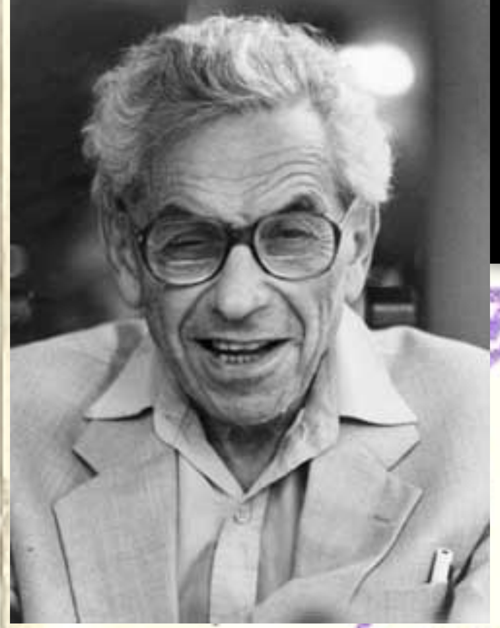
Origins of a small-world idea:

The Bacon number:

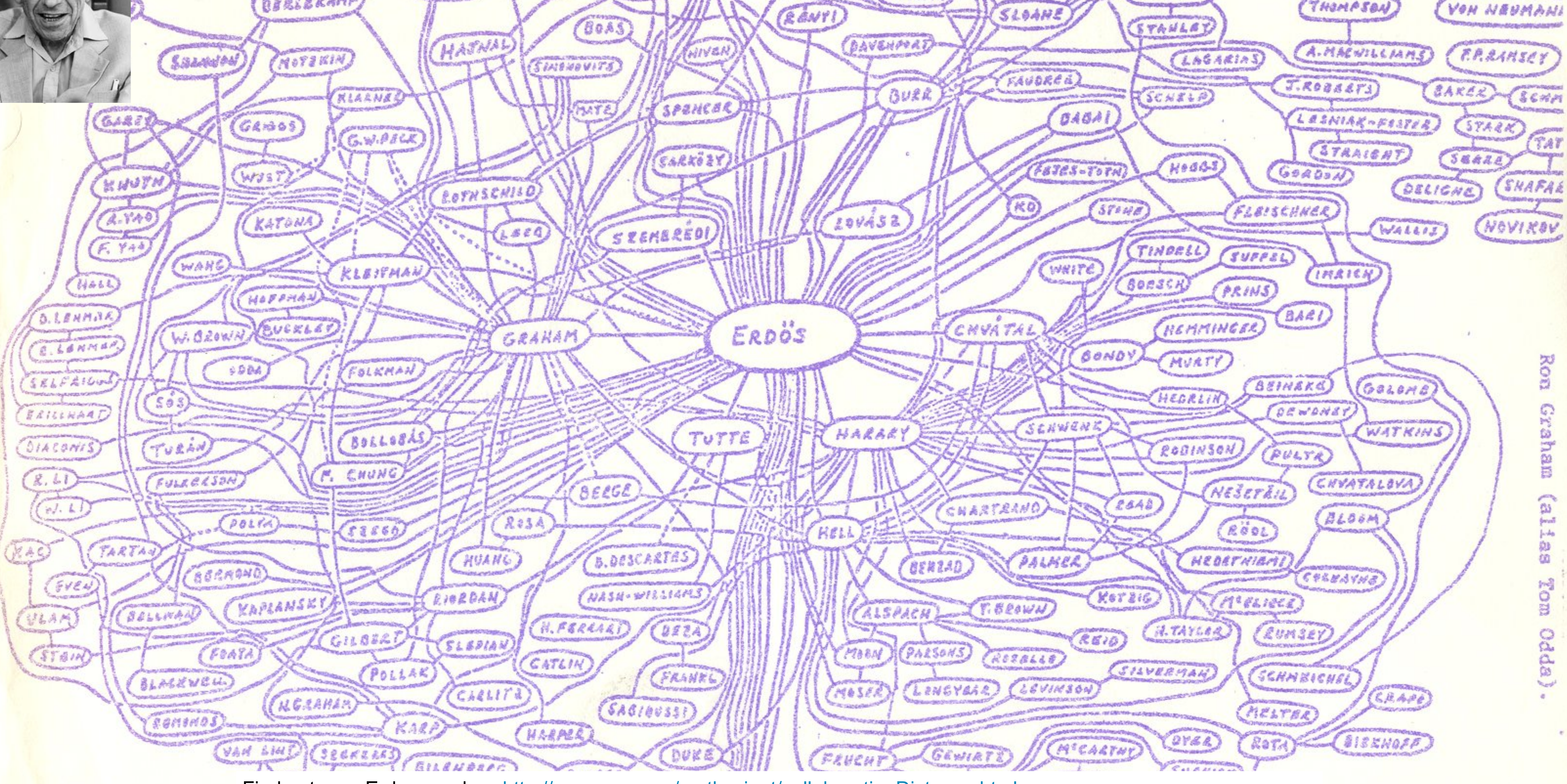
- Create a network of Hollywood actors
- Connect two actors if they co-appeared in the movie
- Bacon number: number of steps to Kevin Bacon

The highest (finite) Bacon number reported is 8

Only approx. 12% of all actors cannot be linked to Bacon (what does this mean about the structure of the actor co-appearance network?)



Erdös numbers are small!
(Ashton's number is 3)



Find out your Erdos number: http://www.ams.org/mathscinet/collaborationDistance.html

Milgram: Further Observations

Criticism

– Starting points and the target were non-random

– There are not many samples (only 64)

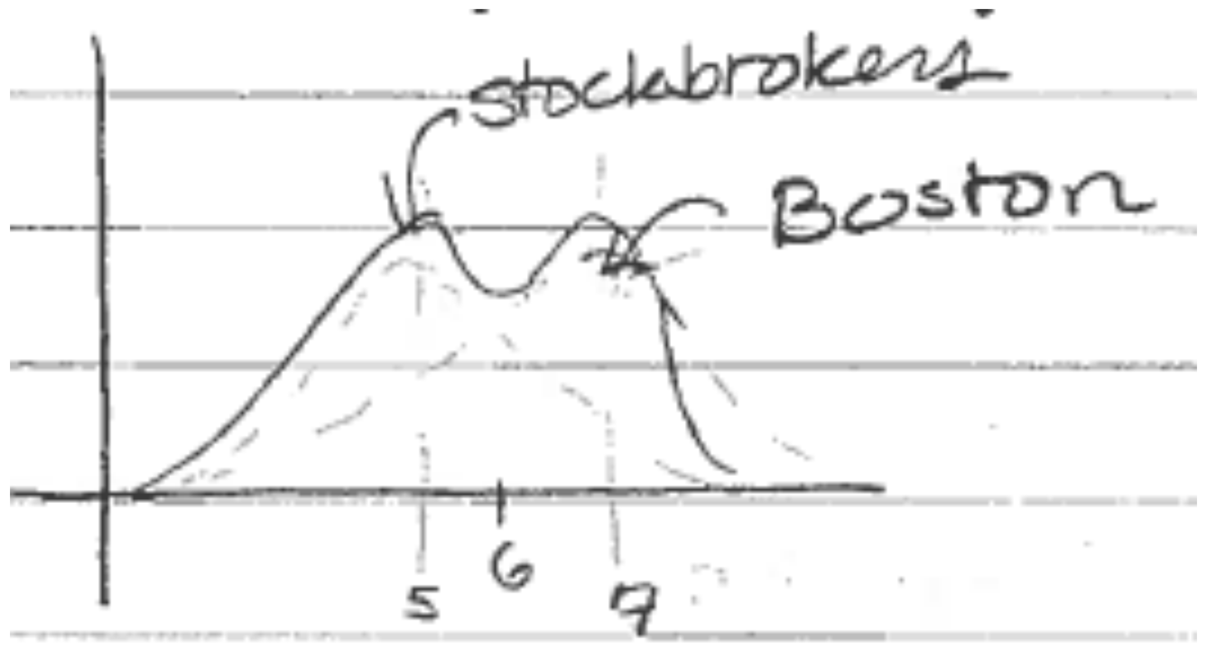
– People refused to participate (25% for Milgram)

Not all searches finished (only 64 out of 300)

– Funneling:

31 of 64 chains passed through 1 of 3 people as their final step → Not all links/nodes are equal

– People might have used extra information resources



Columbia Small-World Study

In 2003 Dodds, Muhamad and Watts performed the experiment using e-mail

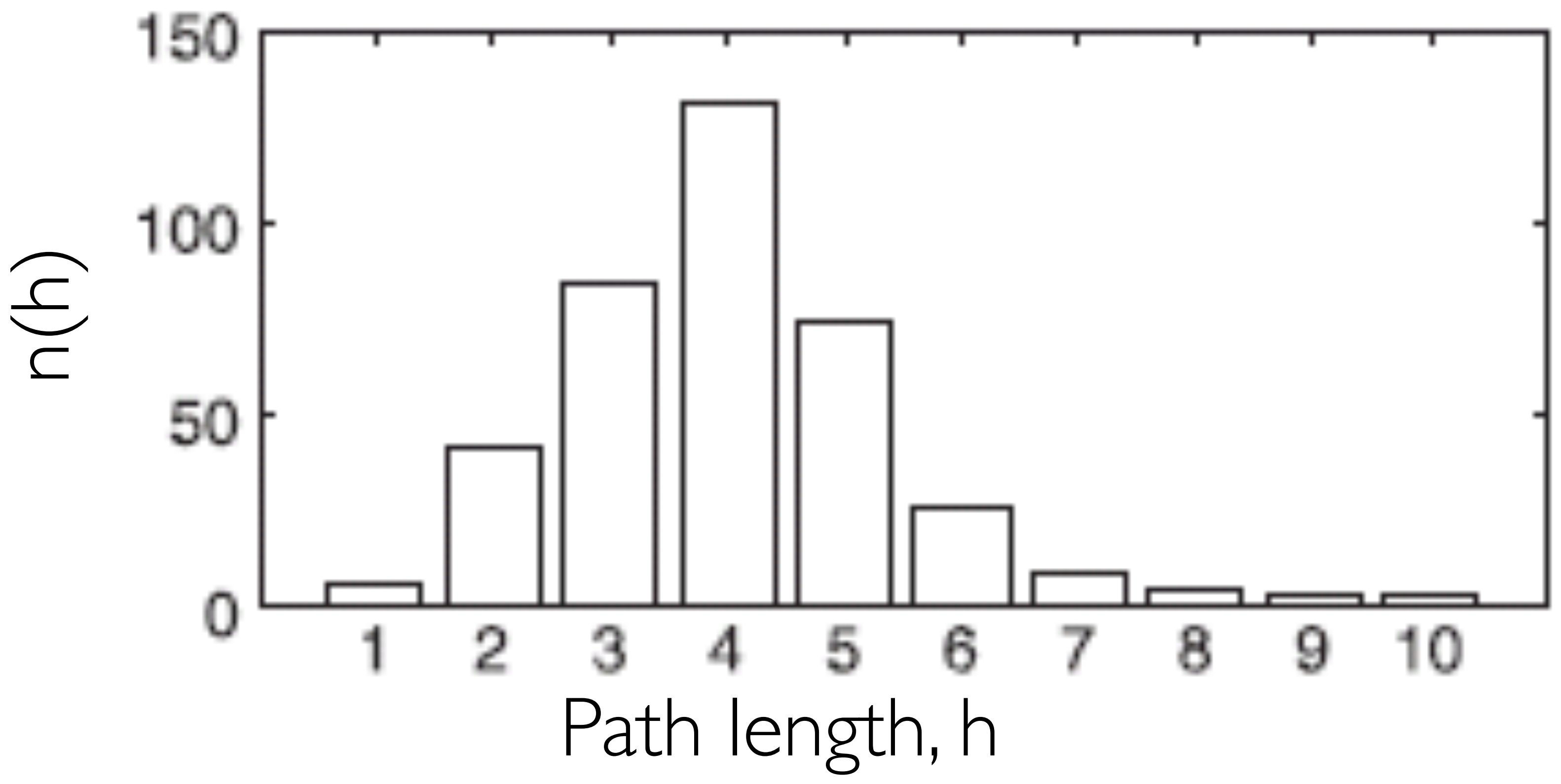
18 diverse targets for the study, including:

- a professor at an Ivy League university,
- an archival inspector in Estonia,
- a technology consultant in India,
- a policeman in Australia,
- a veterinarian in the Norwegian army

Columbia Small-World Study

In 2003 Dodds, Muhamad and Watts performed the experiment using e-mail:

- 18 targets of various backgrounds
- 24,000 first steps (~1,500 per target)
- **65% dropout per step**
- 384 chains completed (1.5%)
- no chain reached the target in Croatia ☹



Avg. chain length = 4.01

Problem: People stop participating

Correction factor:

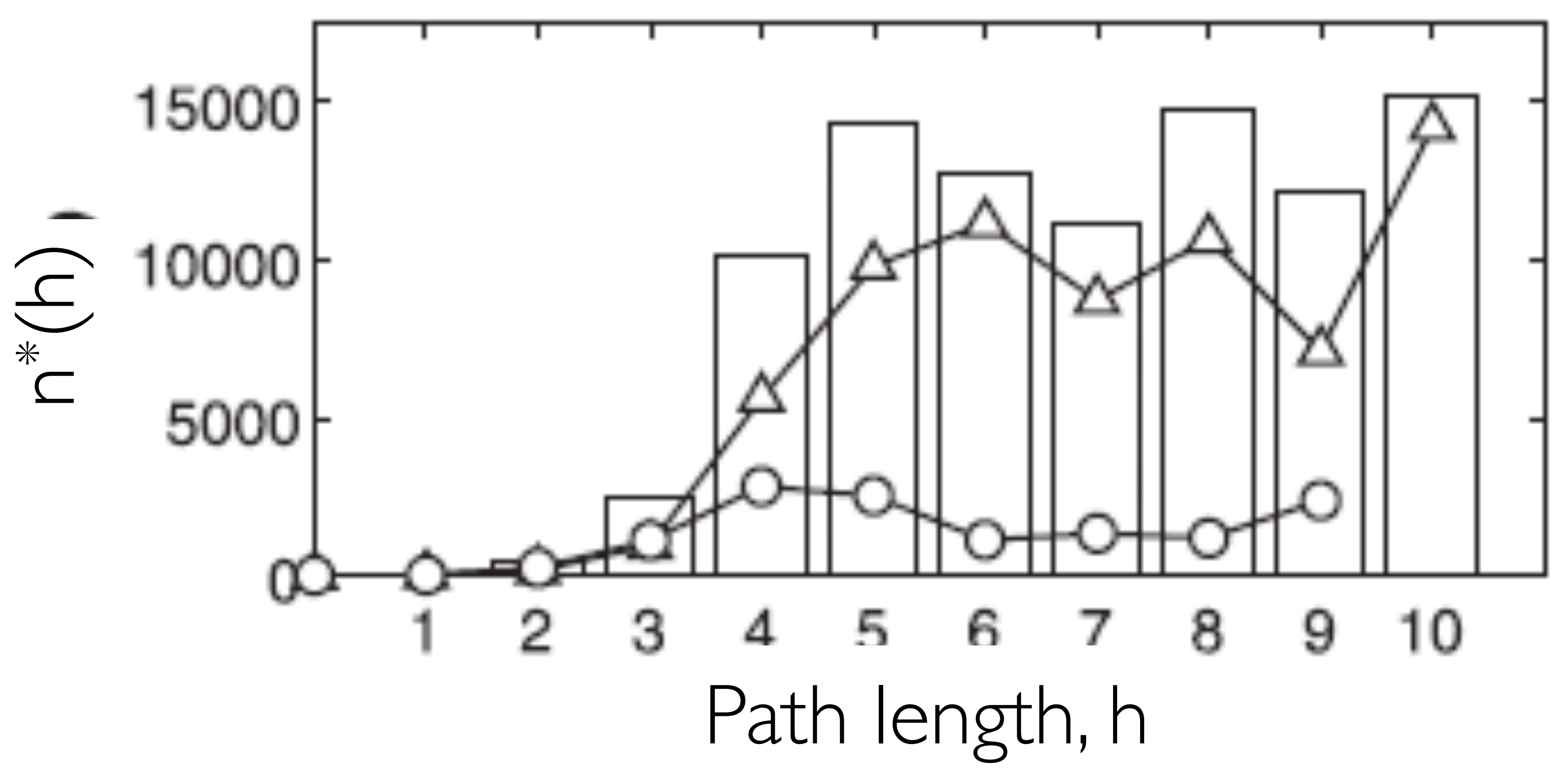
$$n^*(h) = \frac{n(h)}{\prod_{i=0}^{h-1}(1-r_i)}$$

r_i …. drop-out rate at hop i

Small-World in Email Study

After the correction:

Typical path length h = 7



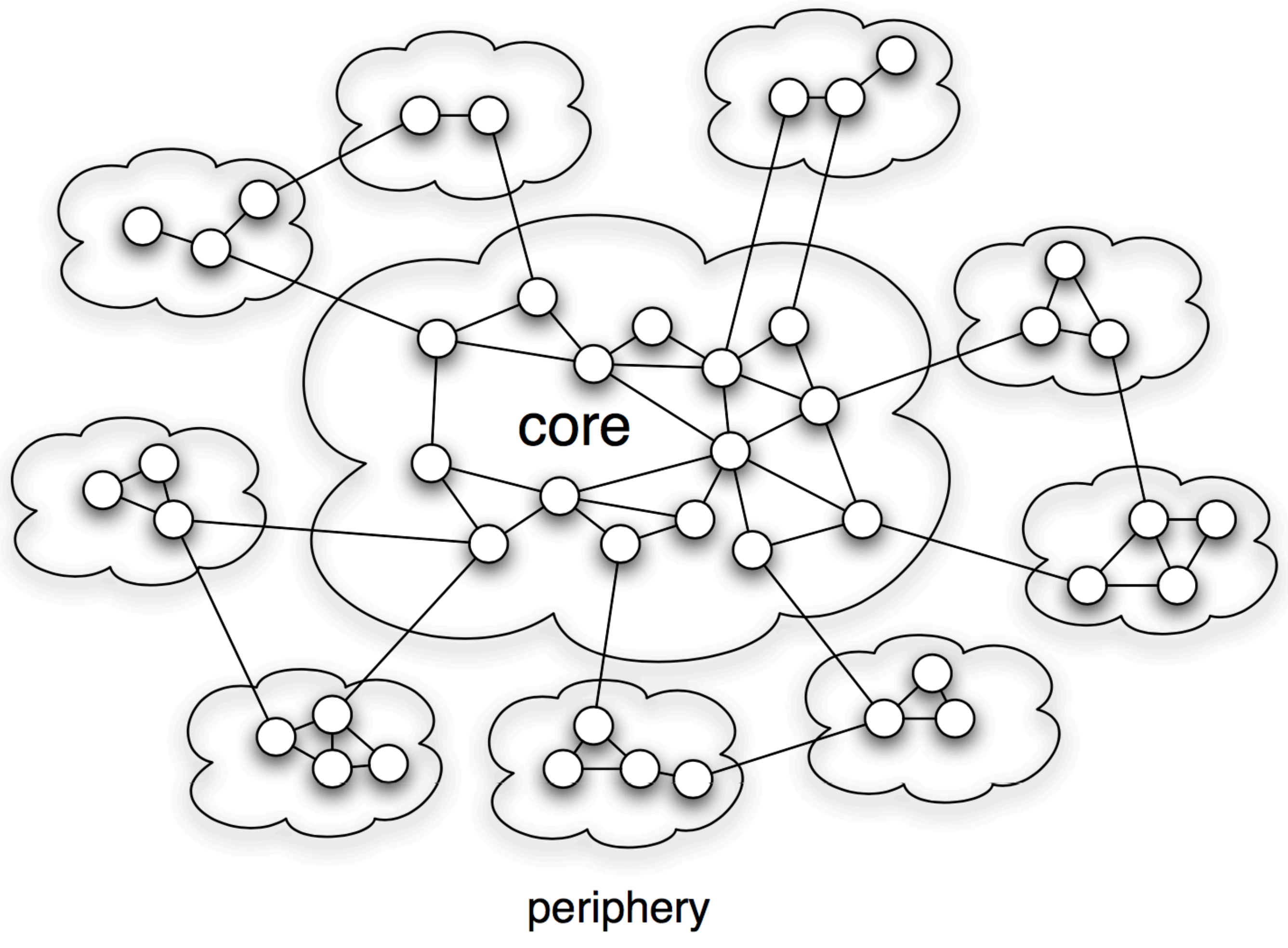
Some not well-understood phenomena in social networks:

Funneling effect: Some target's friends are more likely to be the final step

Conjecture: High reputation/authority

Effects of target's characteristics:
Structurally high-status target easier to find

Conjecture: Core-periphery network structure



Measure on Empirical Data

What are path lengths in real large-scale networks?

Network Analysis Methodology

What are the basic properties properties of real social networks? **Short paths!**

How can we model them? **Path lengths**

Today:

- Milgram's experiment **Letters took 6 hops**
- Measuring path lengths in real-world networks **<— You are here**
- Comparing with a baseline: G_{np} model
- More realistic models: Watts&Strogatz model
- More realistic models: Kleinberg's Decentralized search

Recap: Key Network Properties

Degree distribution: $P(k)$

Clustering coefficient: C

(done in Lecture 2)

Path length: h

(today)

Back to MSN Messenger

MSN Messenger activity in June 2006:

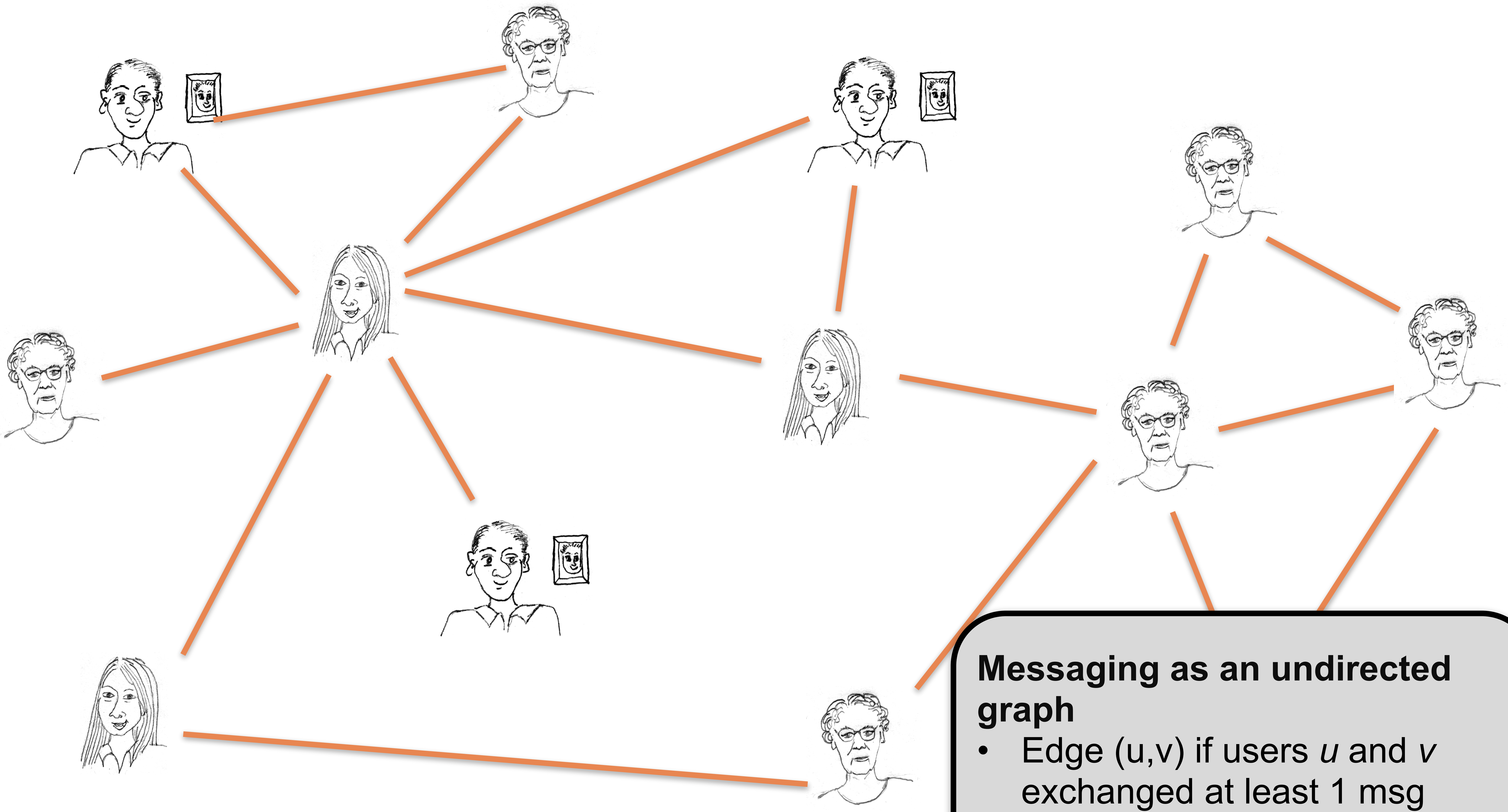
245 million users logged in

180 million users engaged in conversations

More than 30 billion conversations

More than 255 billion exchanged messages

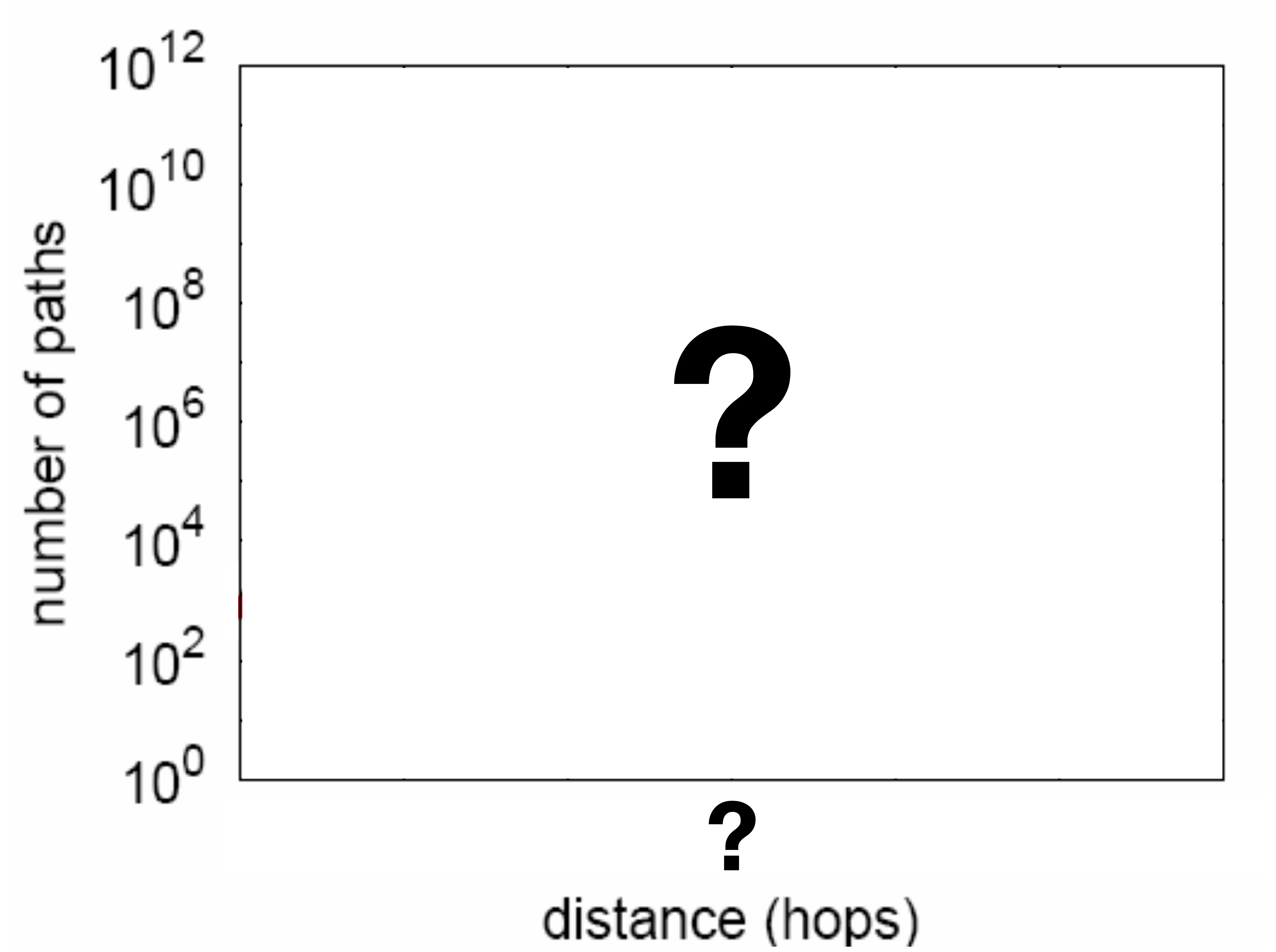
Messaging as a simple graph



Messaging as an undirected graph

- Edge (u,v) if users *u* and *v* exchanged at least 1 msg
- N=180 million people
- E=1.3 billion edges

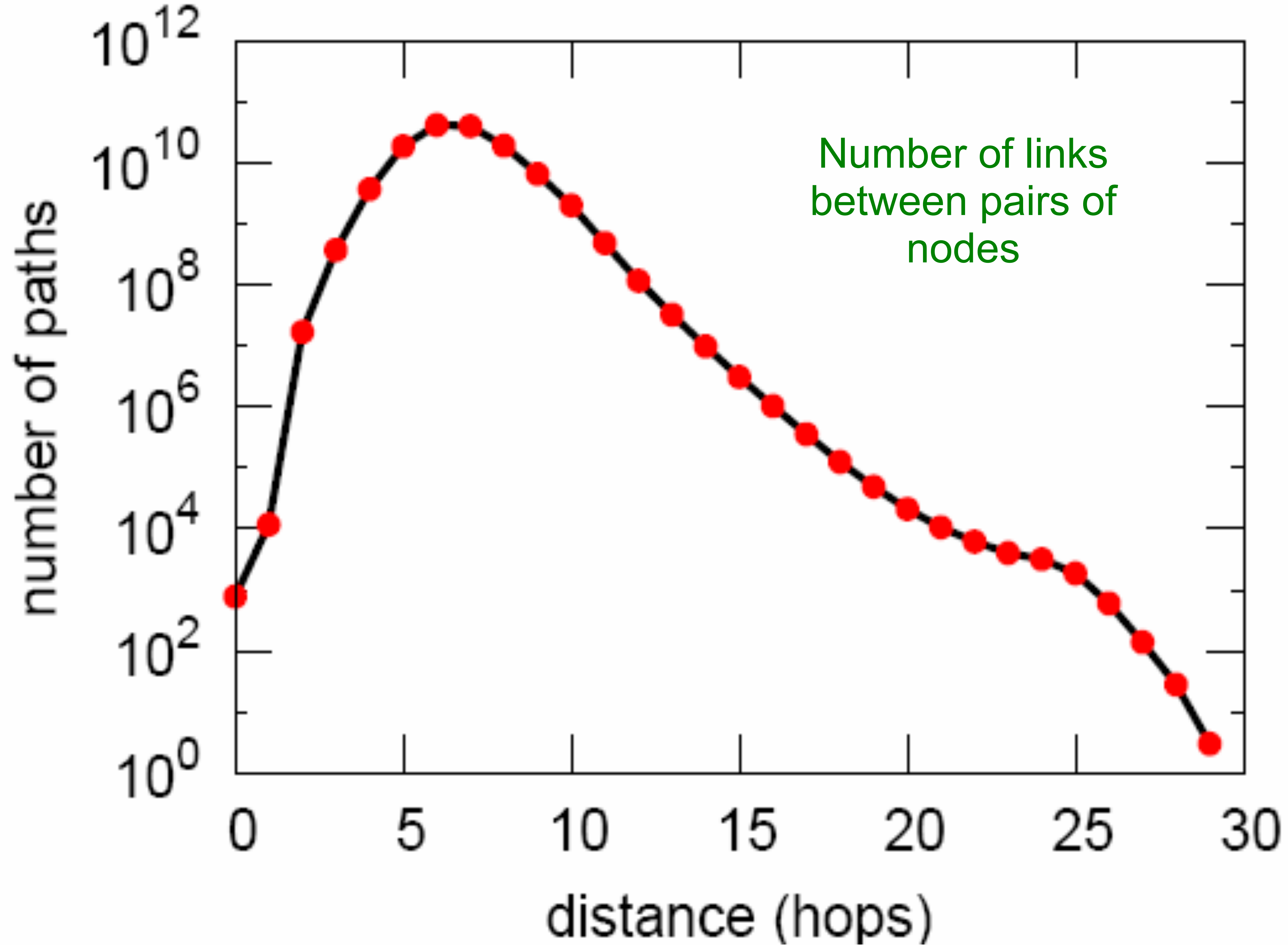
MSN: Path lengths



Number of links between pairs of nodes

How many

MSN: Path lengths



Avg. path length **6.6**
90% of the nodes can be reached in < 8 hops

nodes as we do BFS out of a random node

Steps	#Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

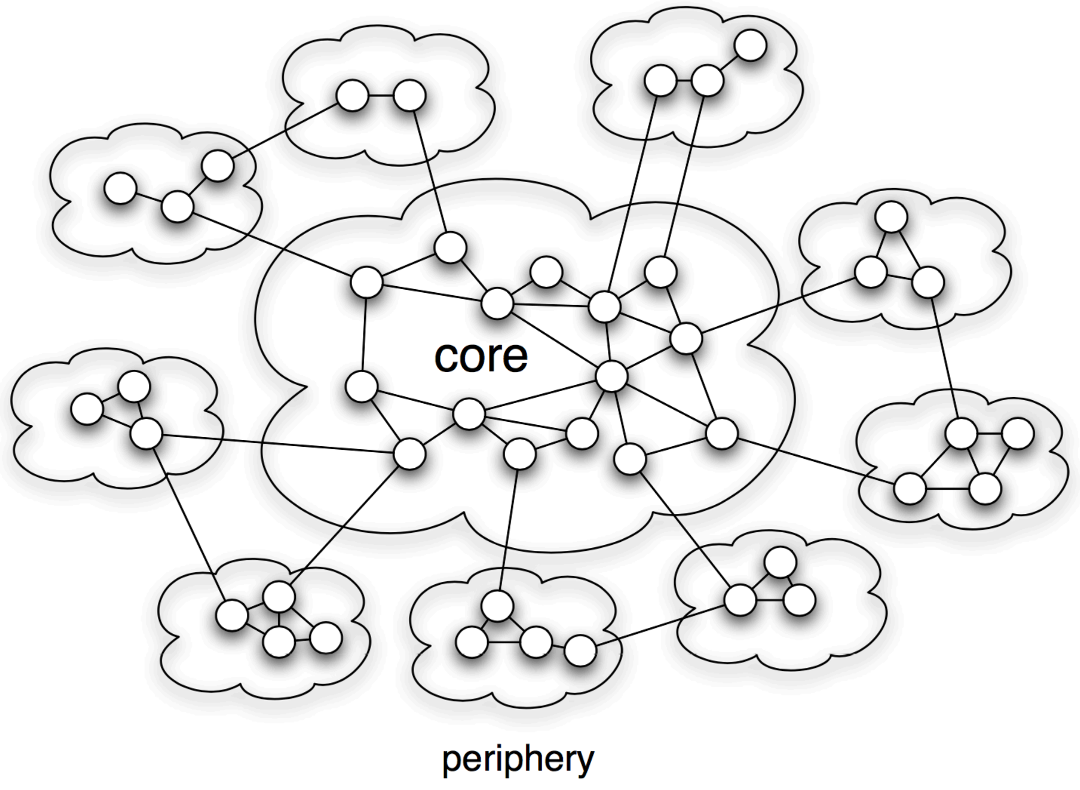
MSN: Key Network Properties

Degree distribution:	*Heavily skewed,* *avg. degree= 14.4*
Clustering coefficient:	*0.11*
Path length:	*6.6*

Short Paths in Empirical Networks

People appear to be surprisingly
well-connected to each other

Do you think this is surprising?



6 Degrees: Should We Be Surprised?

Assume each human is connected to 100 other people

Then:

Step 1: reach 100 people

Step 2: reach 100*100 = 10,000 people

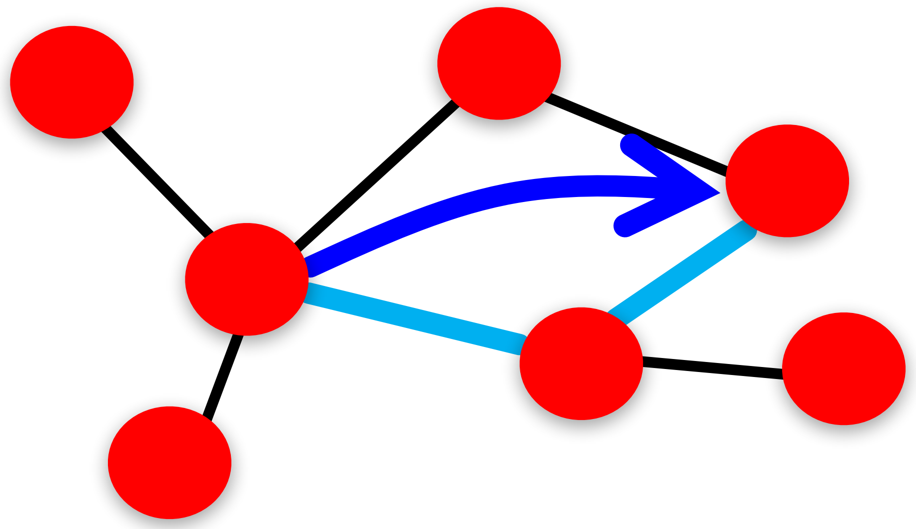
Step 3: reach 100*100*100 = 1,000,000 people

Step 4: reach 100*100*100*100 = 100M people

In 5 steps we can reach 10 billion people

What's wrong here?

Triadic closure: 92% of new FB friendships are to a friend-of-a-friend [Backstom-Leskovec '11]



Back to G_{np}

Erdös-Renyi Random Graphs [Erdös-Renyi, '60]

$G_{n,p}$: undirected graph on n nodes and each edge (u,v) appears i.i.d. with probability p

Simplest random model you can think of

Recap: Key Network Properties

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

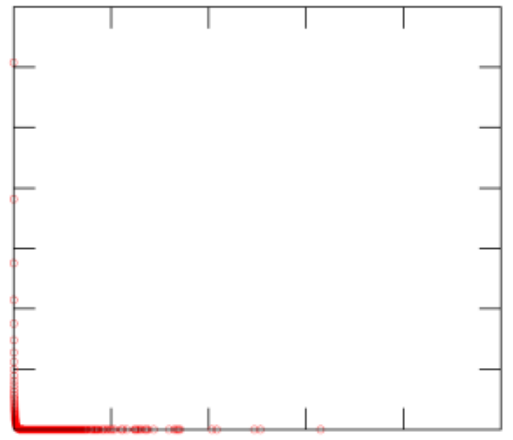
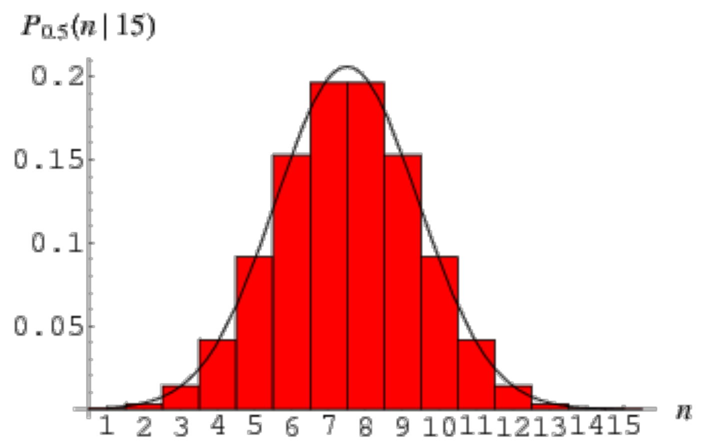
Clustering coefficient: $$C = p = \bar{k}/n$$

(done in Lecture 3)

Path length: $$h \in O(\log n)$$

(today)

MSN vs. G_{np}

	MSN	G_{np}
Degree distribution:		
Clustering coefficient:	0.11	$\bar{k}/n \approx 8 \cdot 10^{-8}$
Path length:	6.6	$h \in O(\log n) \approx 8.2$

Real Networks vs. G_{np}

Are real networks like random graphs?

Average path length: ☺

Clustering Coefficient: ☹

Degree Distribution: ☹

Network Analysis Methodology

What are the basic properties properties of real social networks? **Short paths!**

How can we model them? **Path lengths**

Today:

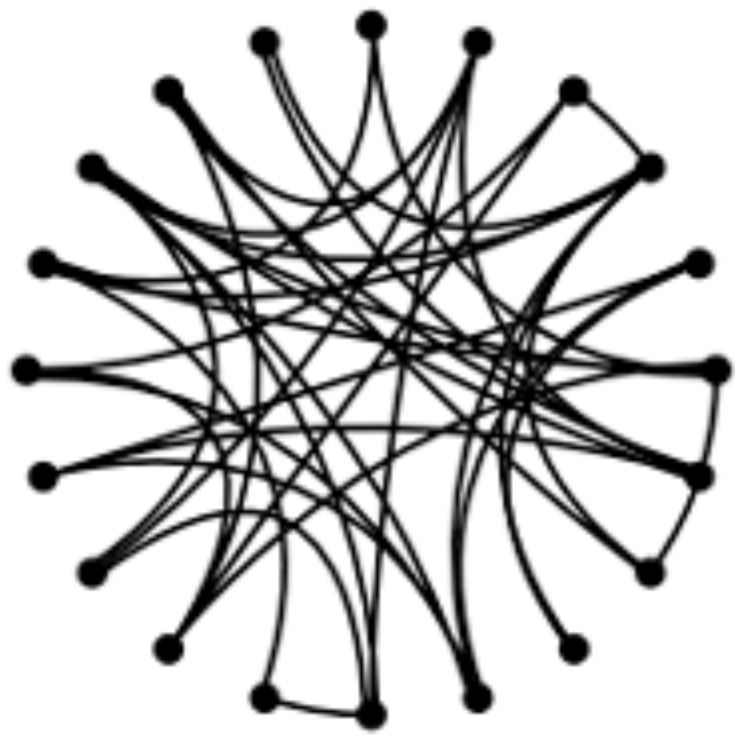
- Milgram's experiment **Letters took 6 hops**
- Measuring path lengths in real-world networks **6.6**
- Comparing with a baseline: G_{np} model ☺ ☹ ☹
- More realistic models: Watts&Strogatz model **<— You are here**
- More realistic models: Kleinberg's Decentralized search

Small World: How?

Can a network with **communities** be a **small world** at the same time?

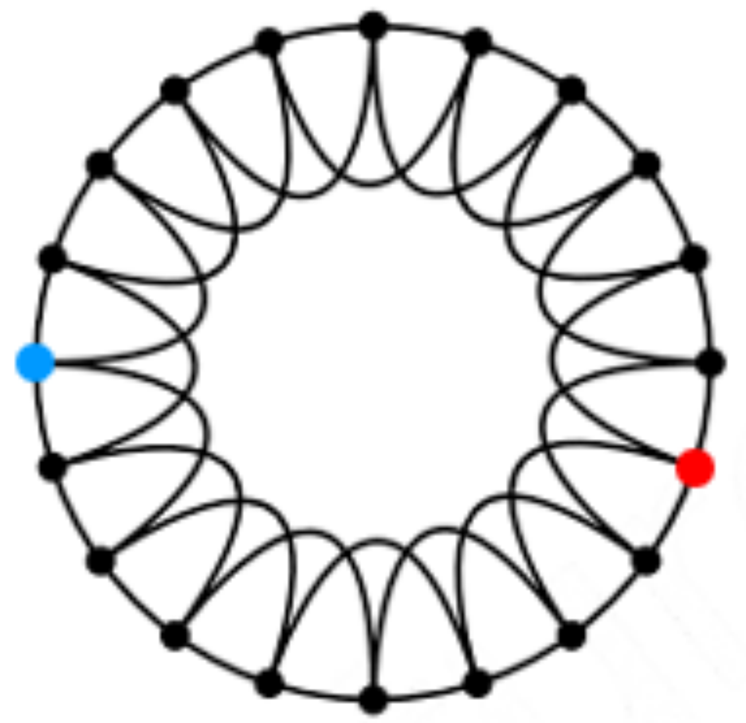
How can we at the same time have **high clustering** and **short paths?**

G_{np}



Low clustering
Short paths

Ring



High clustering
Long paths

Clustering implies edge "locality"

But we need "long-range" edges for short paths

The Small-World Model

Small-world Model [Watts-Strogatz '98]

Two components to the model:

(1) Start with a **low-dimensional regular lattice**

(In this case we use a ring as a lattice)

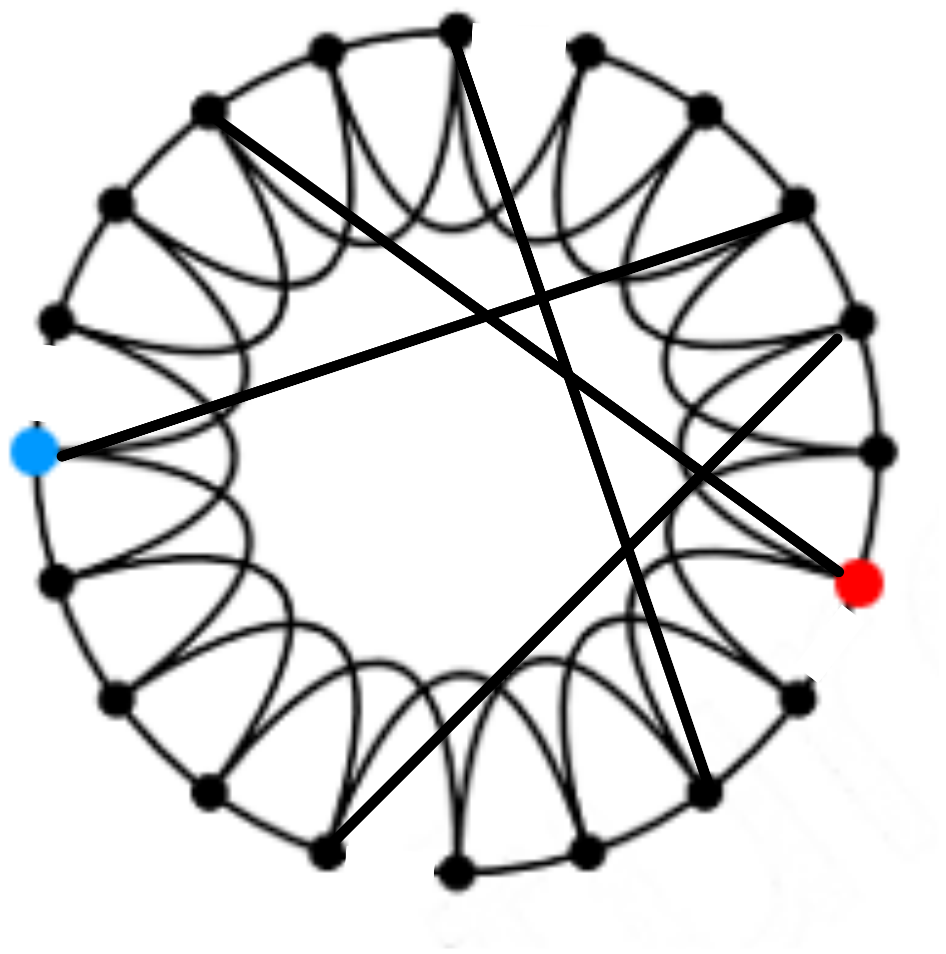
Has high clustering coefficient

Now introduce randomness ("shortcuts")

(2) Rewire:

Add/remove edges to create shortcuts to join remote parts
of the lattice

For each edge with prob. p move the other end to a random node



Watts-Strogatz in 2D

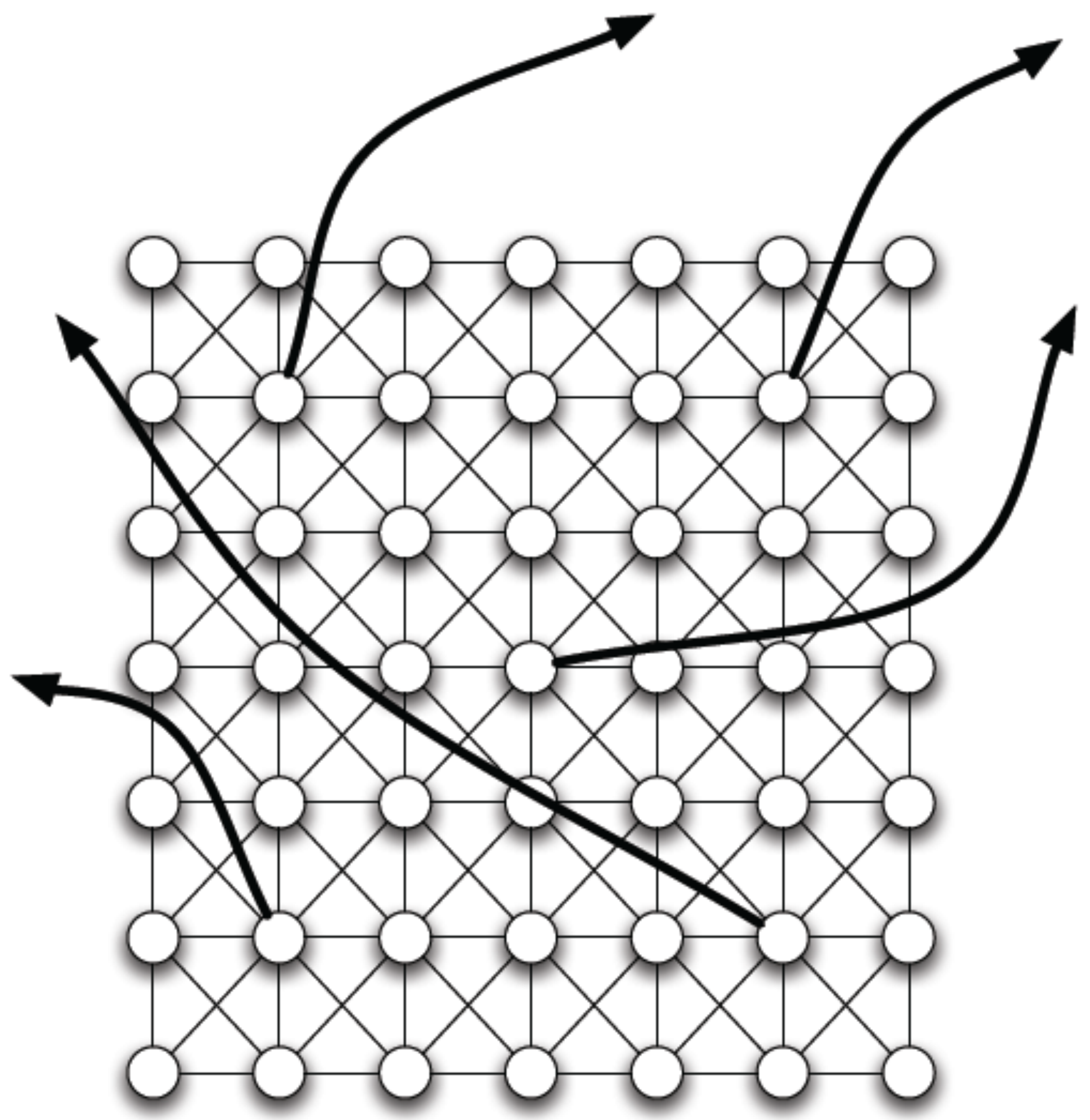
Grid world with some random links

Start with a square grid (two-dimensional square grid)

Two kinds of edges

Link to all other nodes of some radius r

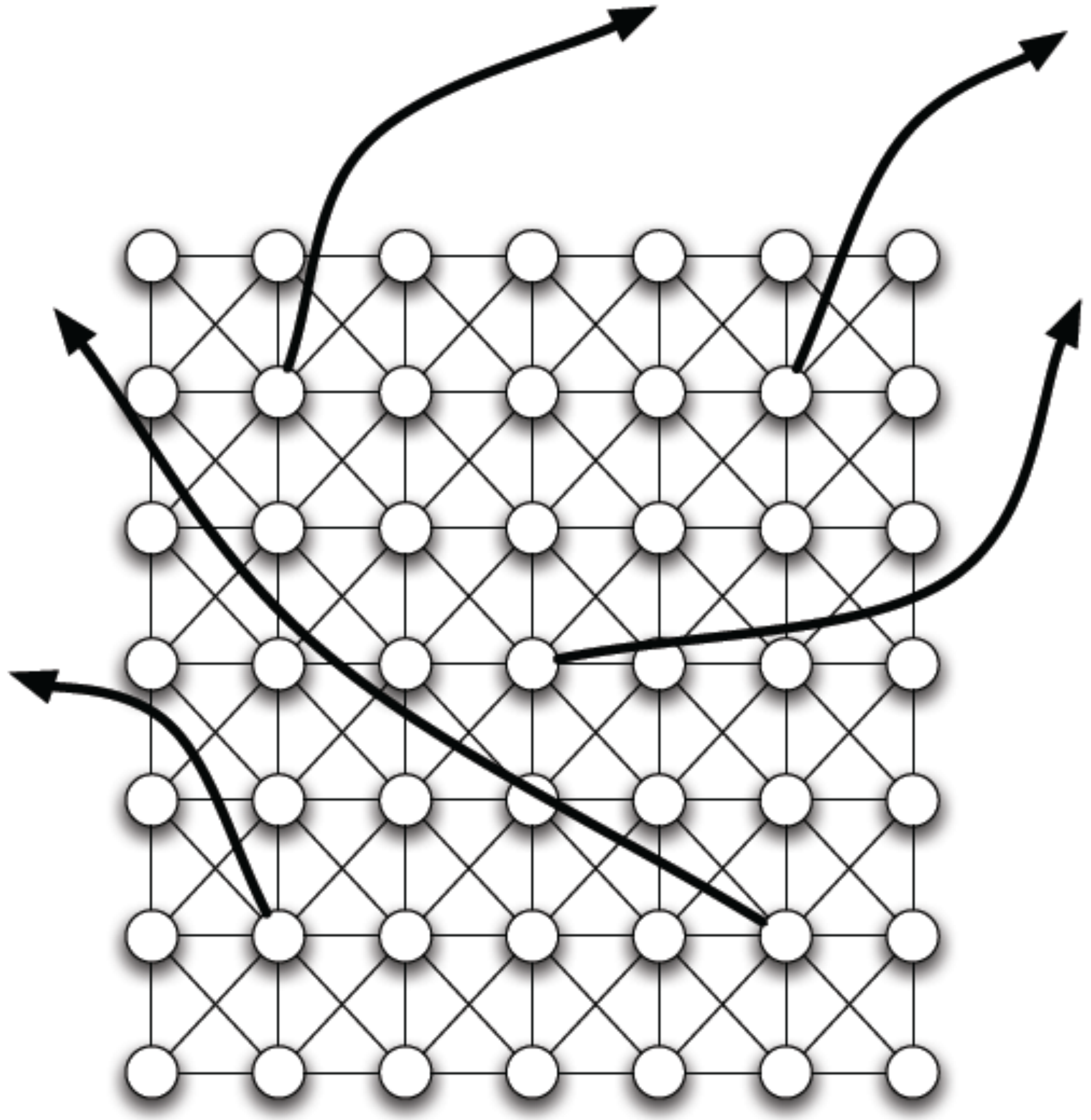
Then add *k* random links per node (like weak ties)



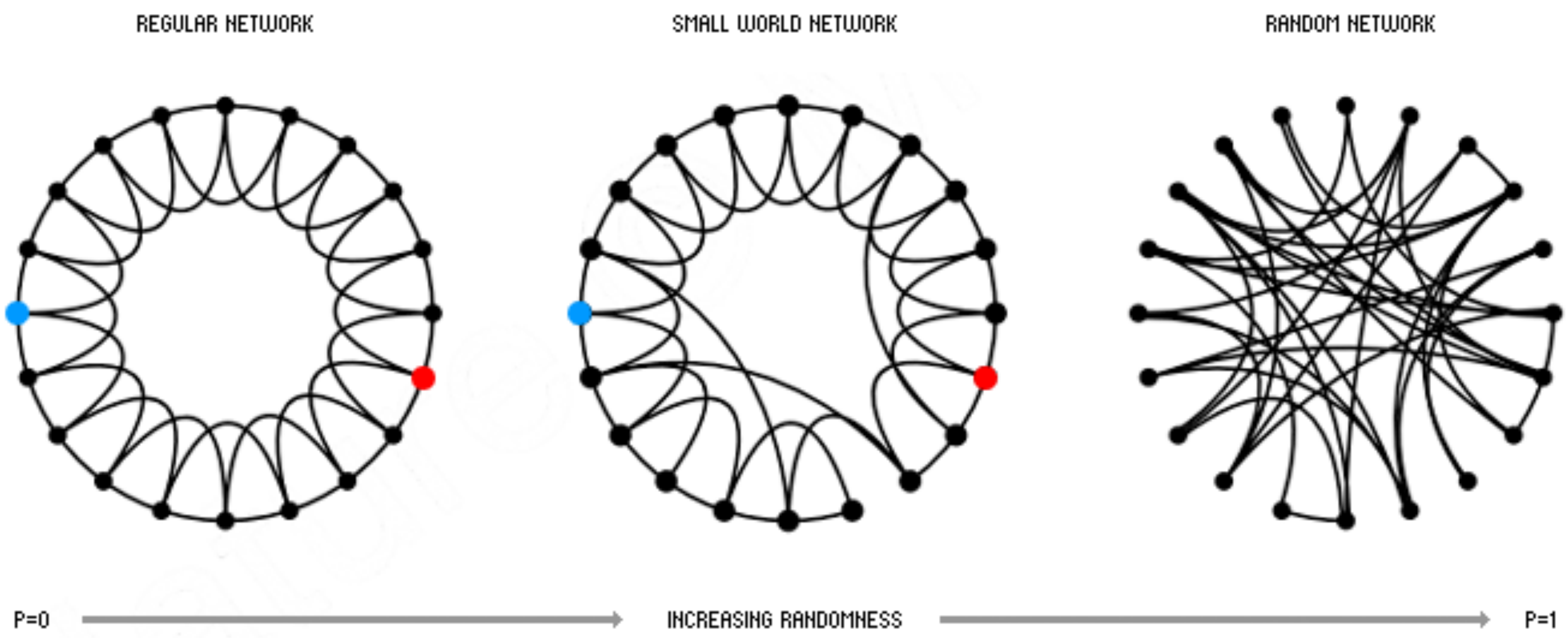
Watts-Strogatz

Lots of clustering, since friends are likely to have friends in common (overlapping neighborhoods)

And short paths! Imagine just following the random links



The Small-World Model



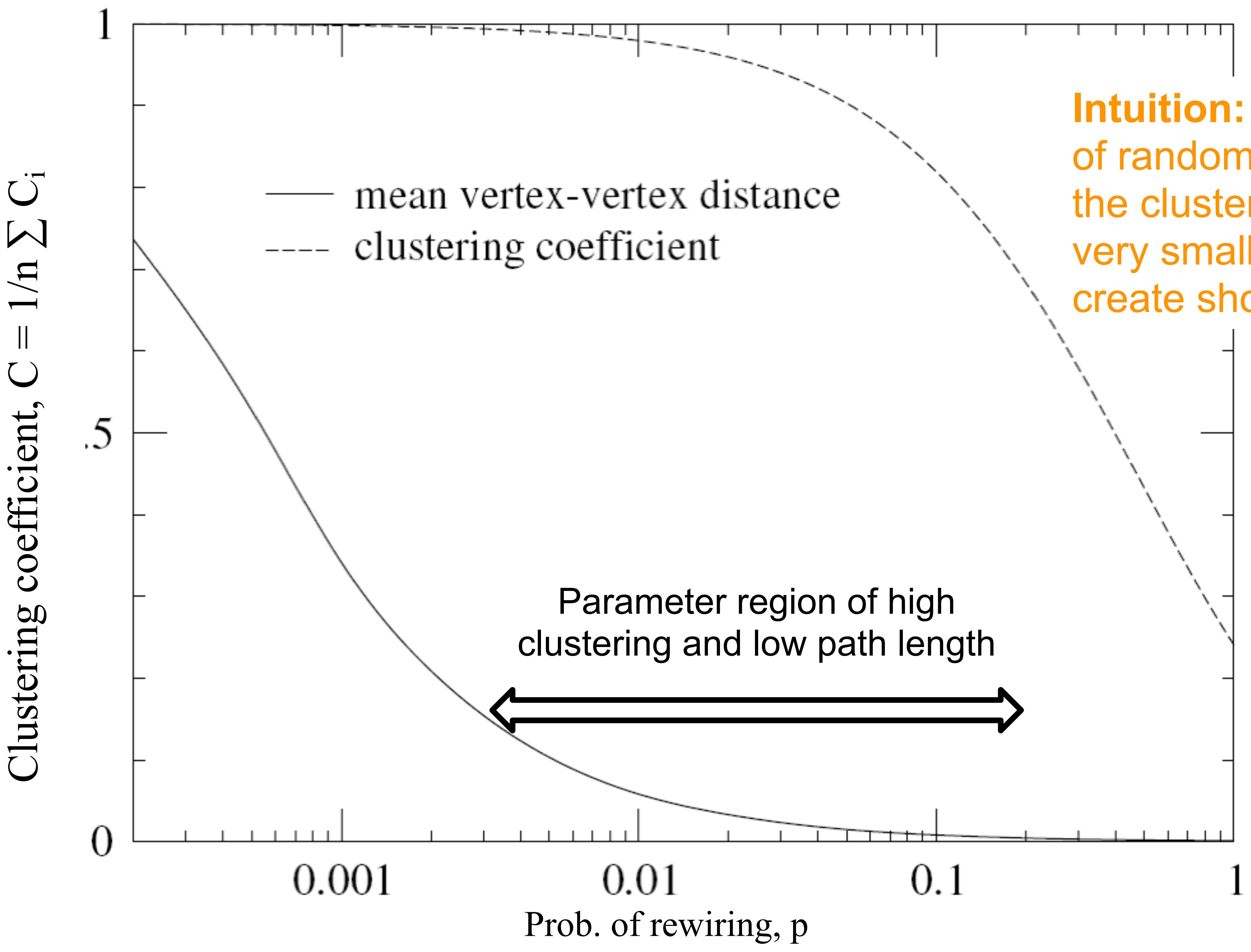
High clustering
High diameter

High clustering
Low diameter

Low clustering
Low diameter

Rewiring allows us to “interpolate” between
a regular lattice and a random graph

The Small-World Model



Intuition: It takes a lot of randomness to ruin the clustering, but a very small amount to create shortcuts.

Small-World: Summary

Could a network with high clustering be at the same time a small world?

Yes! You don't need more than a few random links

The Watts Strogatz Model:

Provides insight on the interplay between clustering and the small-world

Captures the structure of many realistic networks

Accounts for the high clustering of real networks

Does not lead to the correct degree distribution

Network Analysis Methodology

What are the basic properties properties of real social networks? **Short paths!**

How can we model them? **Path lengths**

Today:

- Milgram's experiment **Letters took 6 hops**
- Measuring path lengths in real-world networks **6.6**
- Comparing with a baseline: G_{np} model ☺ ☹ ☹
- More realistic models: Watts&Strogatz model ☺ ☺ ☹
- More realistic models: Kleinberg's Decentralized search

Back to Milgram

Milgram's experiment actually taught us **two** things

- Short paths exist
- But people can also find them!!

No knowledge of the intricate structure of actual social network

- But enough social/geographical/professional markers that people can find short paths anyway

Back to Milgram

To find actual shortest paths, people would have had to send to ALL their contacts

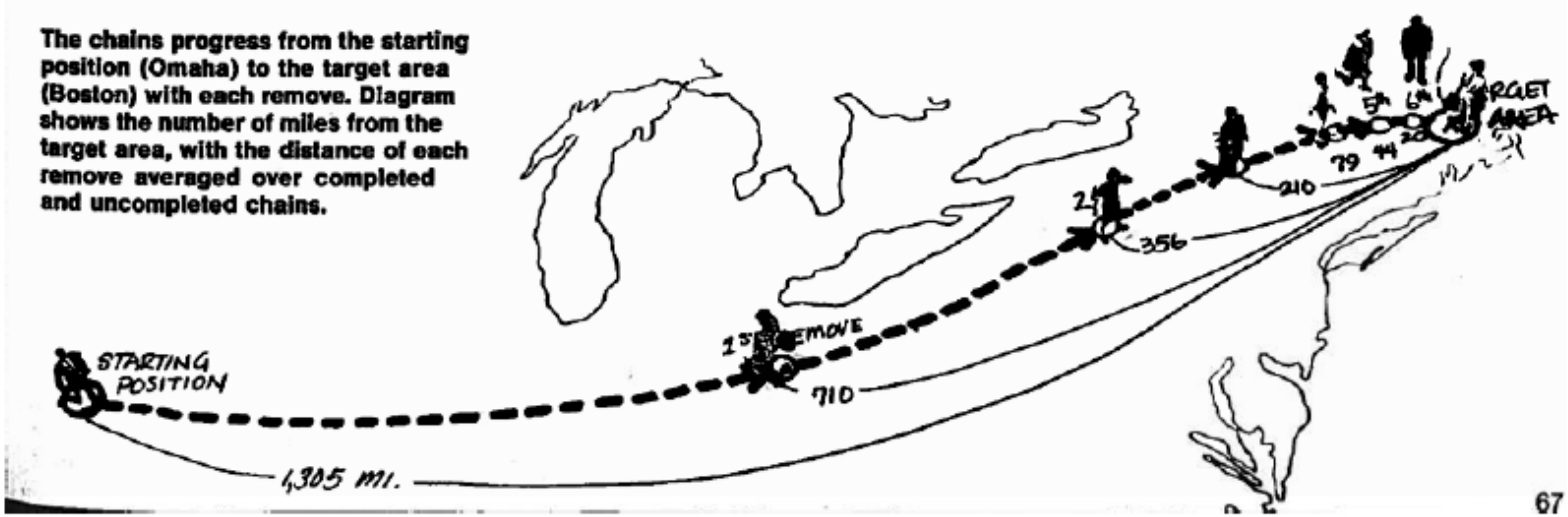
- And hope for 0% attrition
- Simulate BFS

But the actual experiment was much more interesting!

- People had to “tunnel” through the network, doing a kind of decentralized social search
- Could have failed

How to Navigate a Network?

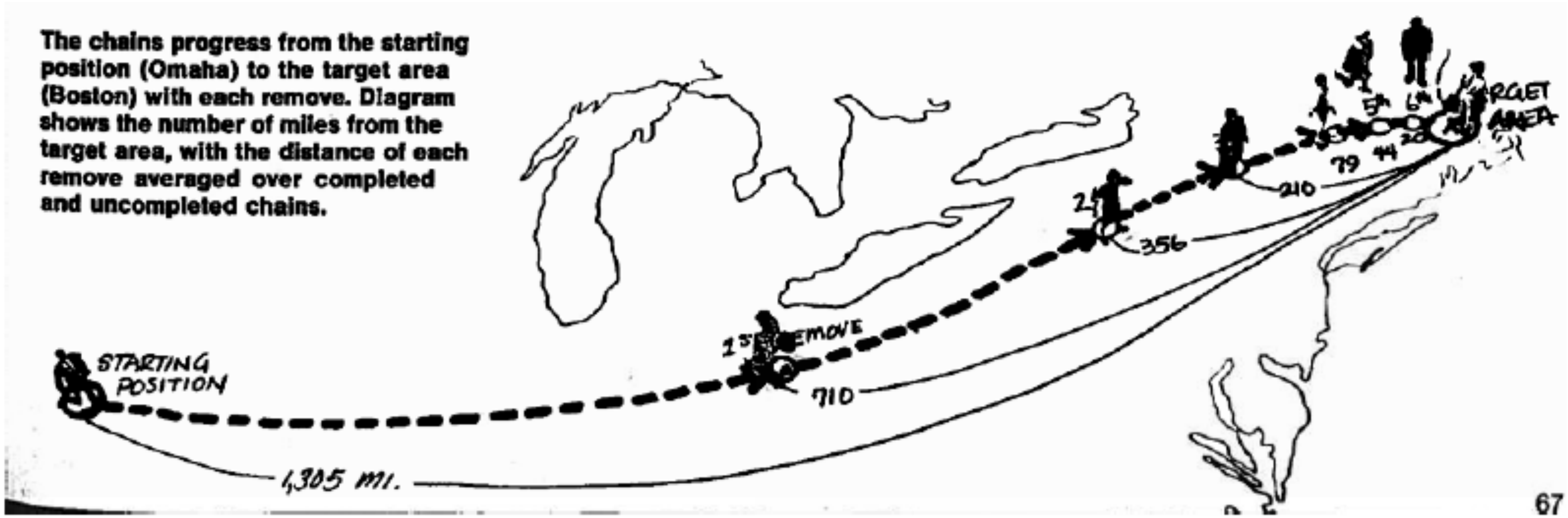
What mechanisms do people use to navigate networks and find the target?



How to Navigate a Network?

"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, 1967

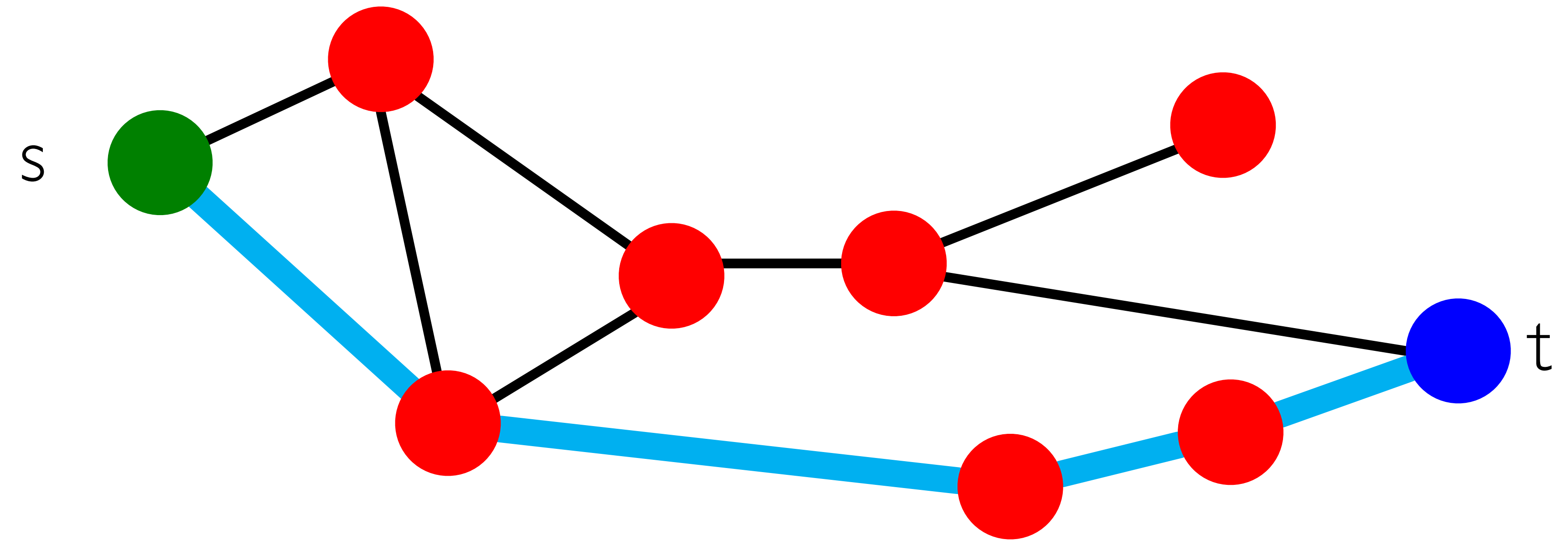


Decentralized Search

The setting:

– Nodes live in a regular lattice, just as in Watts-Strogatz
– Each node has an "address"/location in the grid
– Node ***s*** is trying to route a message to ***t***
– ***s*** only knows locations of its friends and location of the target ***t***
– ***s*** does not know random links of anyone else but itself

We say this kind of search is "decentralized" because no one has complete (i.e. "centralized") knowledge of the network



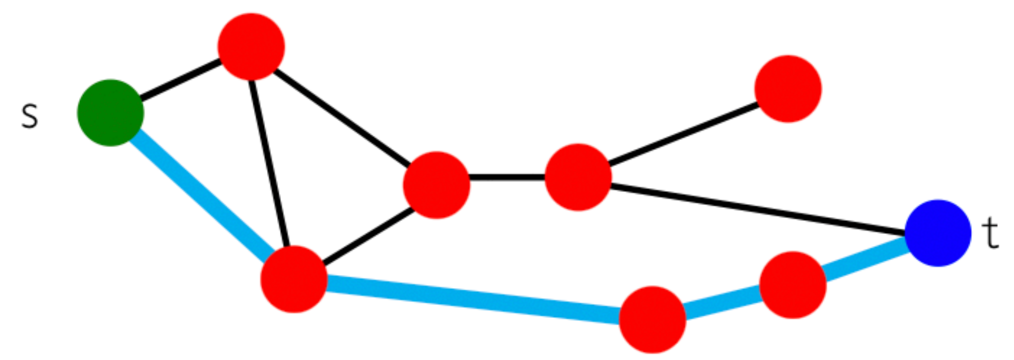
Decentralized Search

The setting:

– Nodes live in a regular lattice, just as in Watts-Strogatz
– Each node has an "address"/location in the grid
– Node ***s*** is trying to route a message to ***t***
– ***s*** only knows locations of its friends and location of the target ***t***
– ***s*** does not know random links of anyone else but itself

Geographic Navigation: nodes will act greedily with respect to geography: always pass the message to their neighbour who is geographically closest to ***t*** (what else can they do?)

Search time T: Number of steps it takes to reach ***t***



What is success?

We know these graphs have diameter O(log n), so paths are logarithmic in shortest-path length

We will say a graph is **searchable** if the decentralised search time T is polynomial in the path lengths

But it's **not searchable** if T is exponential in the path lengths

Searchable

Search time T:

$$O((\log n)^{\beta})$$

Not searchable

Search time T:

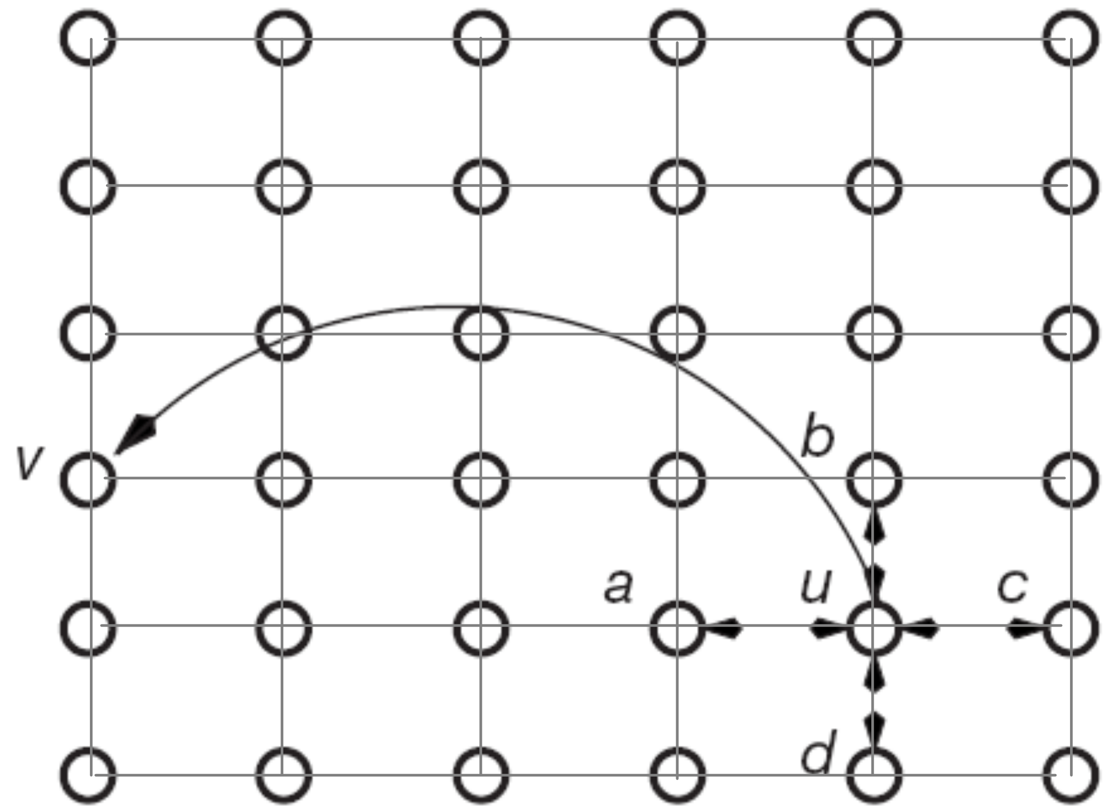
$$O(n^{\alpha})$$

Navigation in Watts-Strogatz

Model: 2-dim grid where each node has 1 random edge

This is a small-world!

(Small-world = diameter O(log n))



Fact: A decentralized search algorithm in Watts-Strogatz model needs $n^{2/3}$ steps to reach t in expectation

Even though paths of ***O(log n)*** steps exist!

Note: All our calculations are asymptotic, i.e., we are interested in what happens as $n \to \infty$

Overview of the Results

Not searchable

Search time T:

$$O(n^{\alpha})$$

Watts-Strogatz

$$O(n^{\frac{2}{3}})$$

Erdös–Rényi (G_{np})

$$O(n)$$

Searchable

Search time T:

$$O((\log n)^{\beta})$$

Next: Kleinberg's model

$$O((\log n)^{2})$$

Navigable Small-World Graph?

Watts-Strogatz graphs are
not searchable

How do we make a searchable
small-world graph?

Navigable Small-World Graph?

Watts-Strogatz graphs are
not searchable

How do we make a searchable small-world graph?

Intuition:

Our long range links are **not random**

They follow **geography**

Saul Steinberg, “View of the World from 9th Avenue”

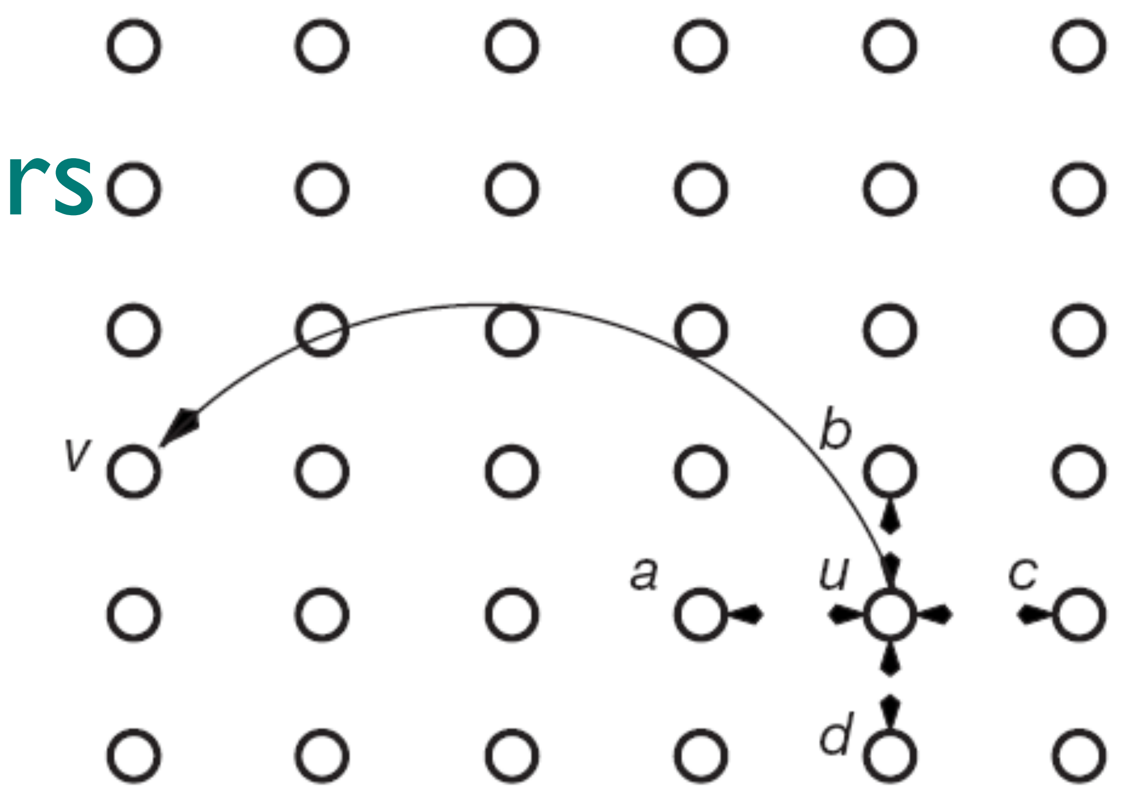
Kleinberg's Model

Kleinberg's Model [Kleinberg, Nature '01]

Nodes still live in a grid, and know their neighbours

Each node has one random "long-range" link

Key difference: the link isn't uniformly at random anymore, it follows geography



Prob. of long link to node *v*:

$$P(u \to v) \sim d(u,v)^{-\alpha}$$

$d(u,v)$ … grid distance between *u* and *v* (address distance, not shortest path)

α … tunable parameter ≥ 0

Kleinberg's Model

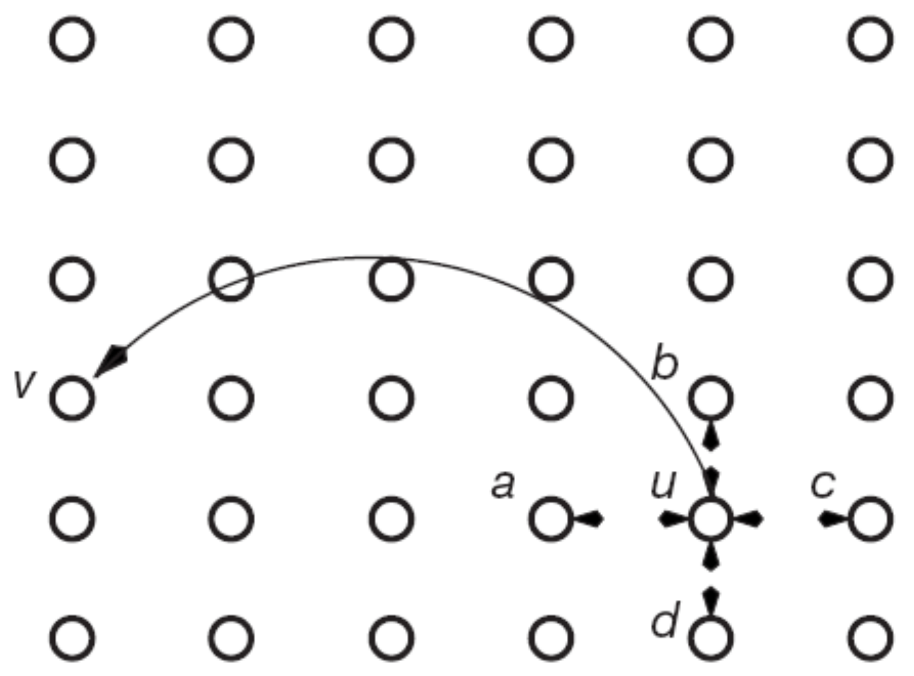
Kleinberg's Model [Kleinberg, Nature '01]

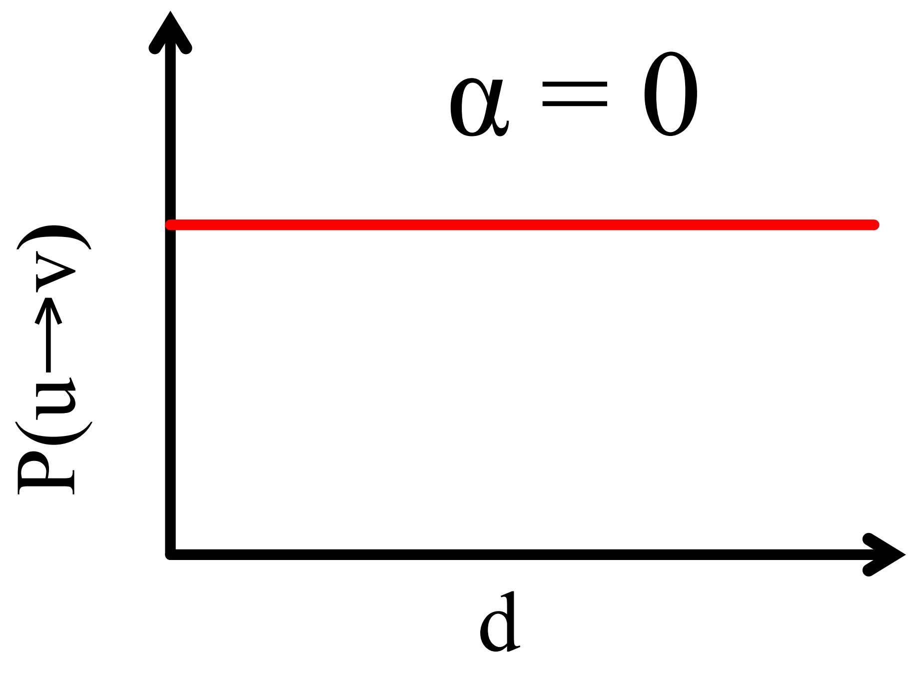
Prob. of long link to node *v*:

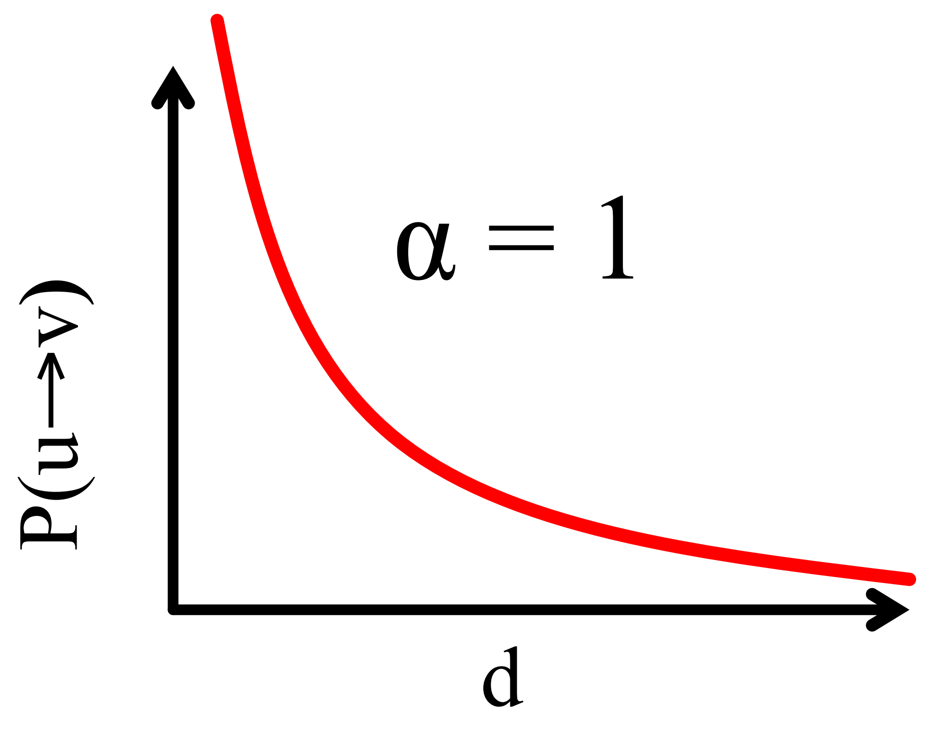
$$P(u \to v) \sim d(u,v)^{-\alpha}$$

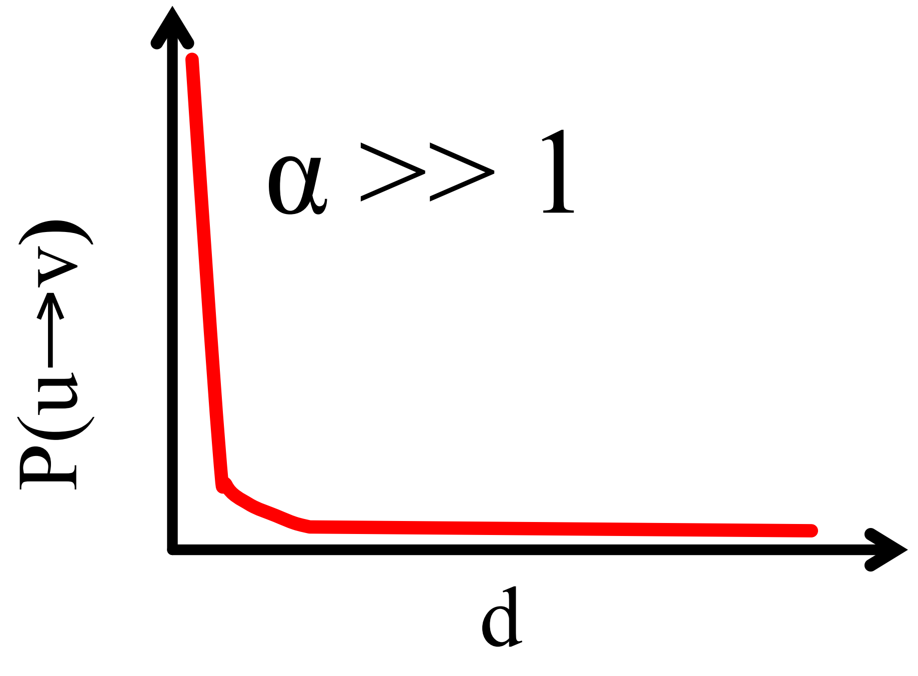
$d(u,v)$ … grid distance between *u* and *v* (address distance, not shortest path)

α … tunable parameter ≥ 0







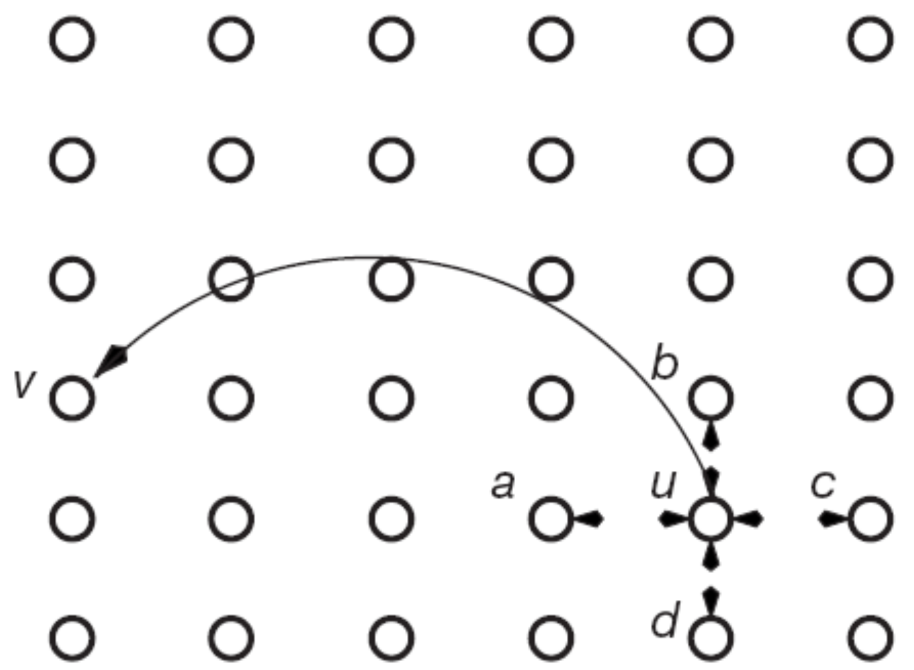


Kleinberg's Model

Kleinberg's Model [Kleinberg, Nature '01]

Prob. of long link to node *v*:

$$P(u \to v) \sim d(u,v)^{-\alpha}$$



Express as a probability by dividing by the proper normalizing constant:

$$P(u \to v) = \frac{d(u,v)^{-\alpha}}{\sum_{w \neq u} d(u,w)^{-\alpha}}$$

This just ensures the probabilities sum to 1

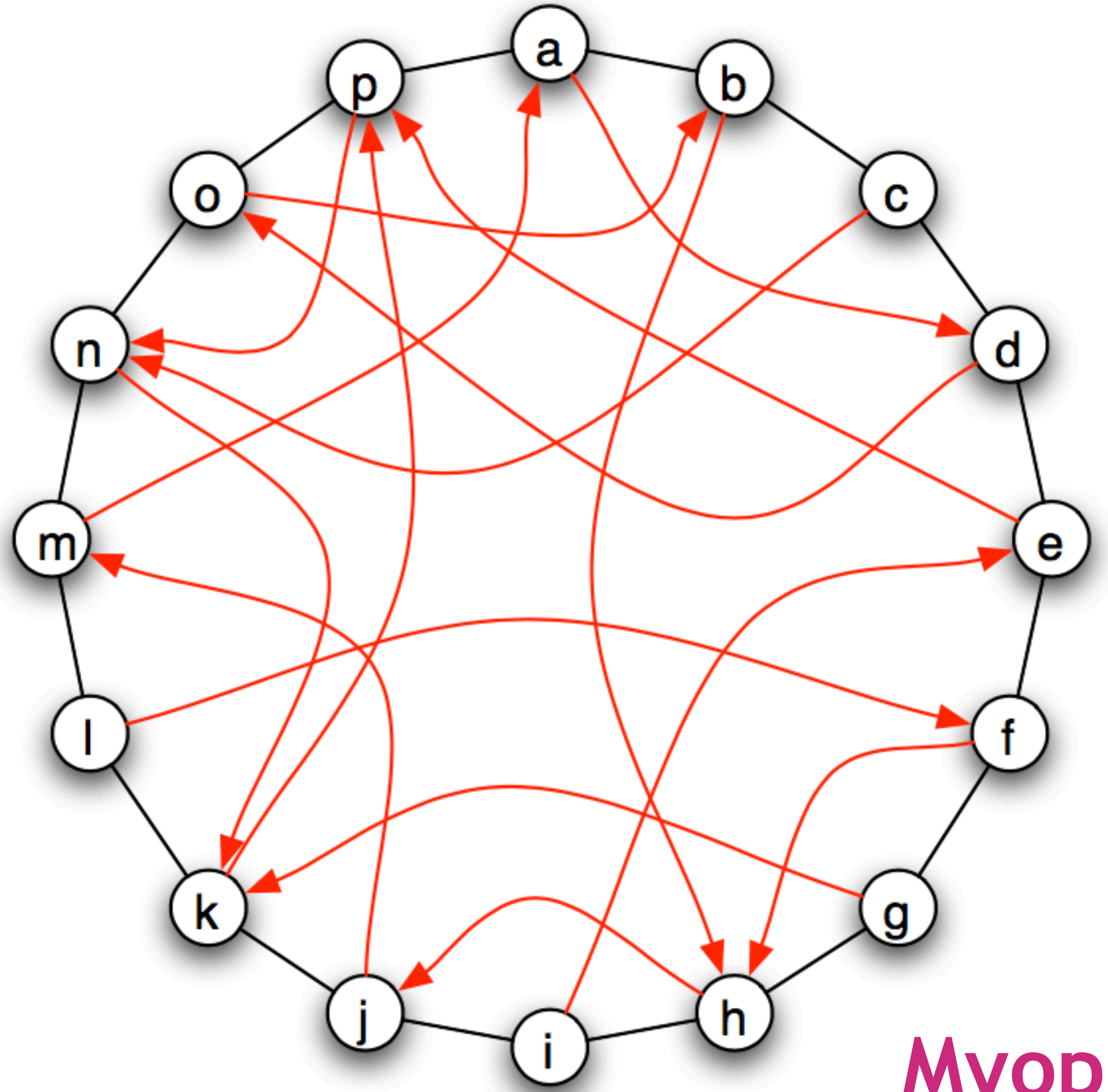
Kleinberg's Model in 1-Dimension

We will analyze the 1-dimensional case

Nodes use the greedy strategy ("myopic search"): at each step, pass to contact geographically closest to the target

Kleinberg's Model in 1-Dimension

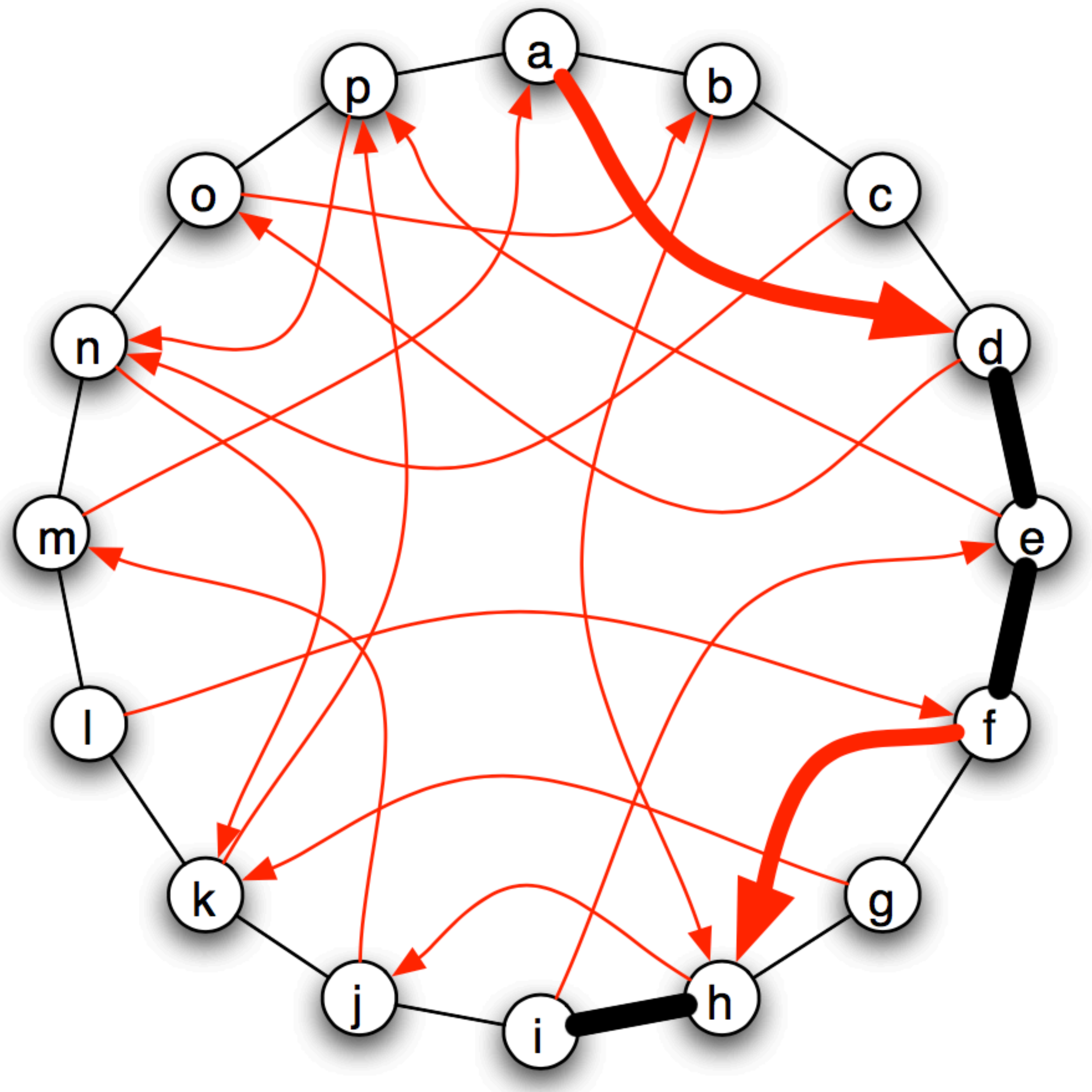
We will analyze the 1-dimensional case



Myopic path a->i?

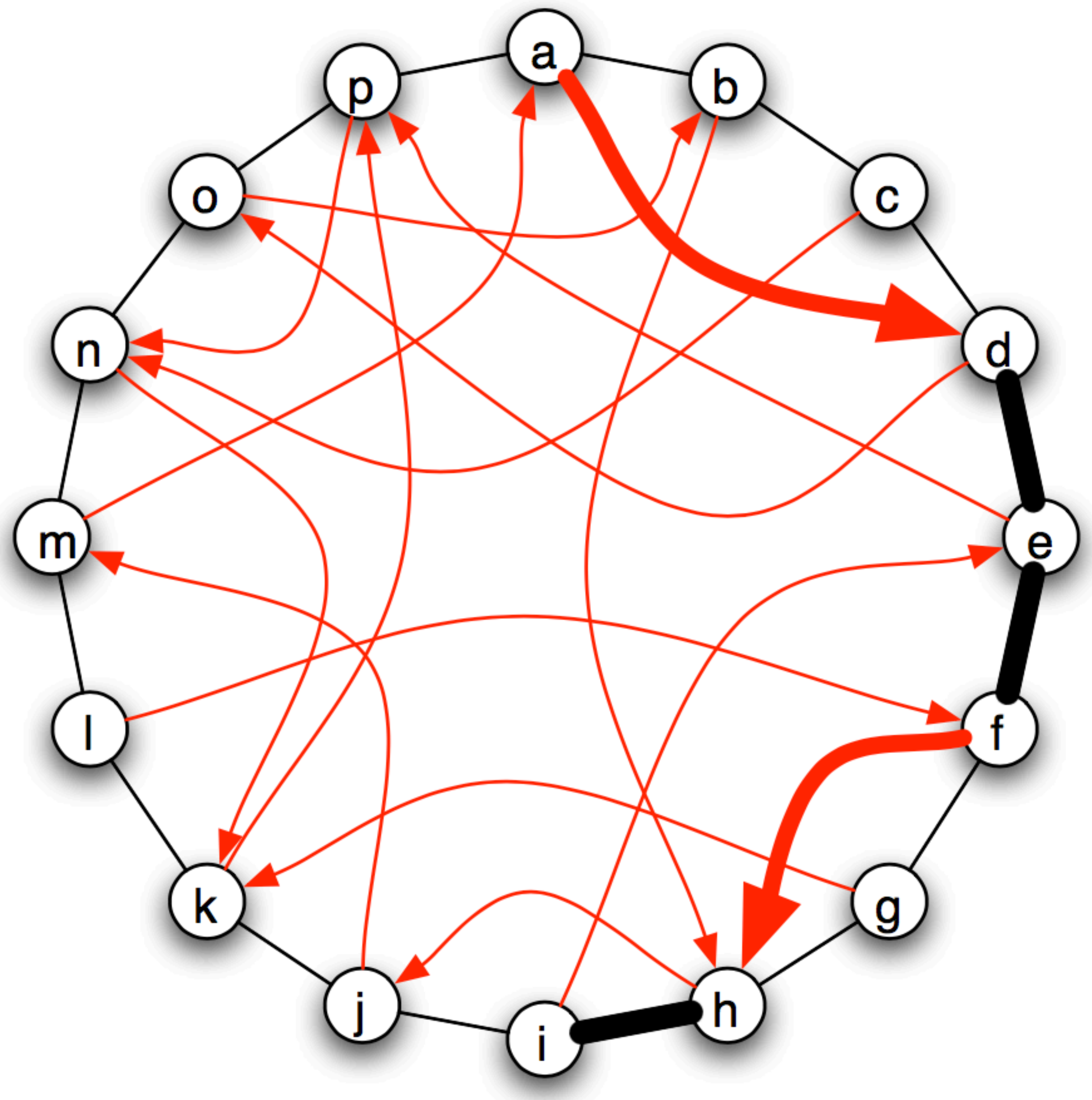
Kleinberg's Model in 1-Dimension

We will analyze the 1-dimensional case



Kleinberg’s Model in 1-Dimension

Not the shortest path!



Kleinberg's Model in 1-Dimension

We now have a completely well-defined probabilistic question:

– Start with a ring where each node its connected to its two neighbours
– Add one random link per node according to geography
– Choose random start s and random target t

What is expected path length of myopic search?

Kleinberg's Model in 1-Dimension

We analyze 1-dimensional case:

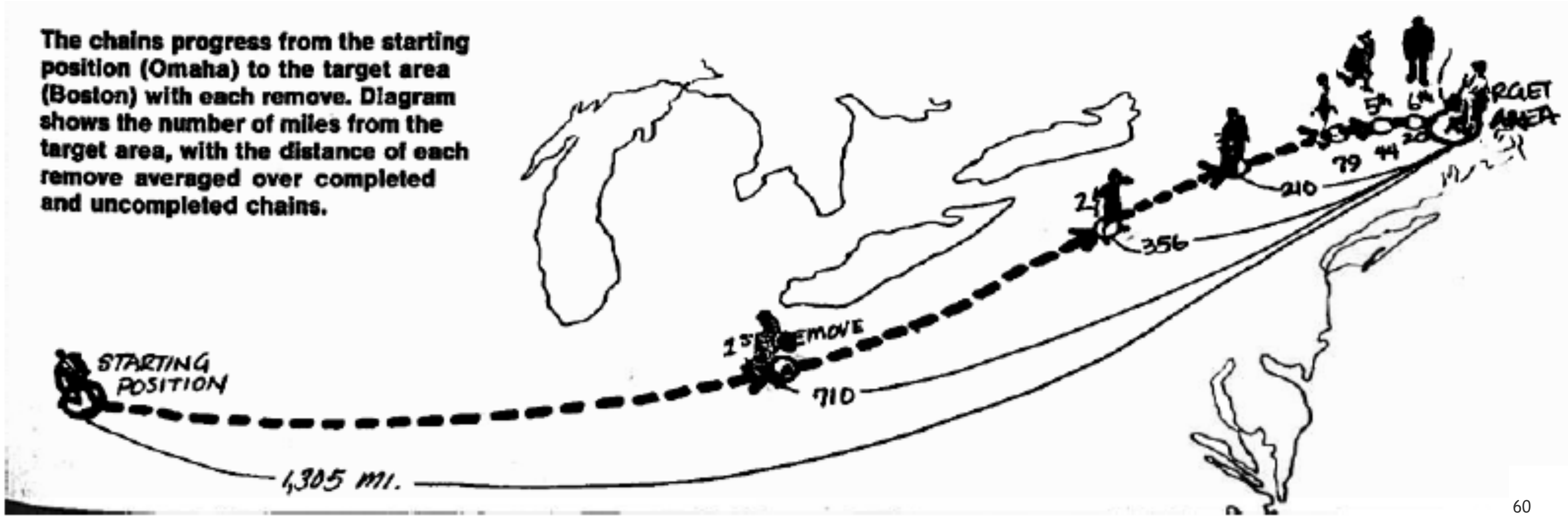
Claim: For $\alpha = 1$ we can get from *s* to *t* in O(log(n)²) steps in expectation

$$P(u \to v) \sim d(u,v)^{-\alpha} = 1/d(u,v)$$

Proof strategy:

Argue it takes O(log n) to halve the distance

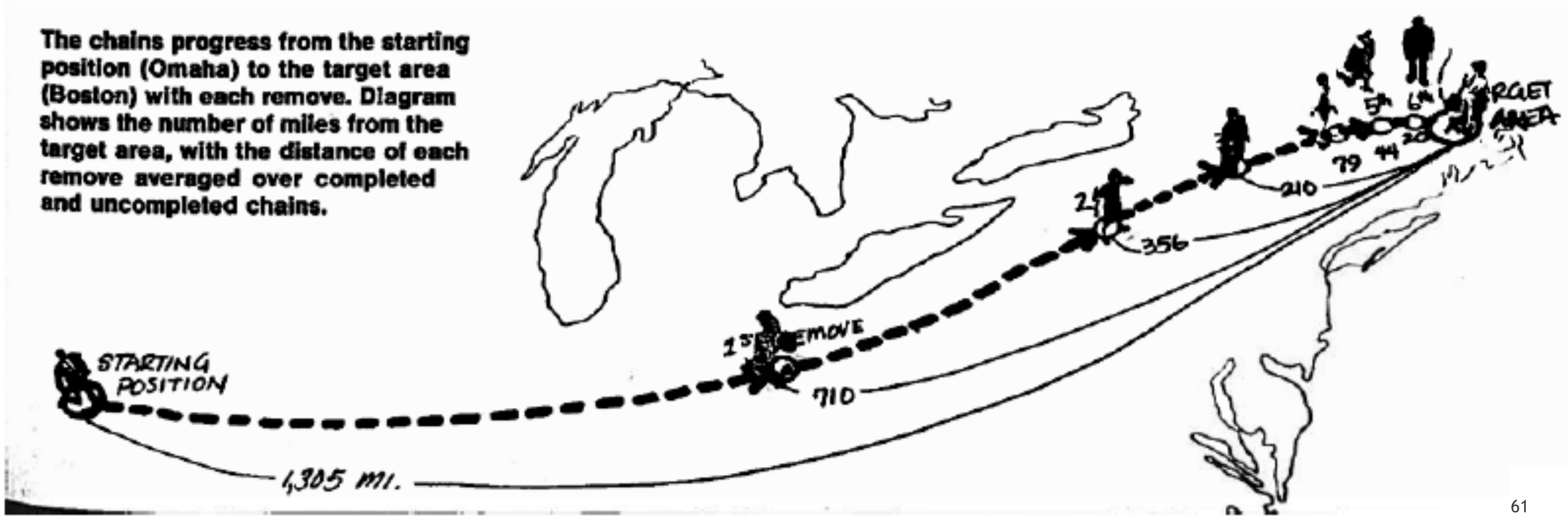
O(log n) halving steps to get to target



Kleinberg’s Model in 1-Dimension

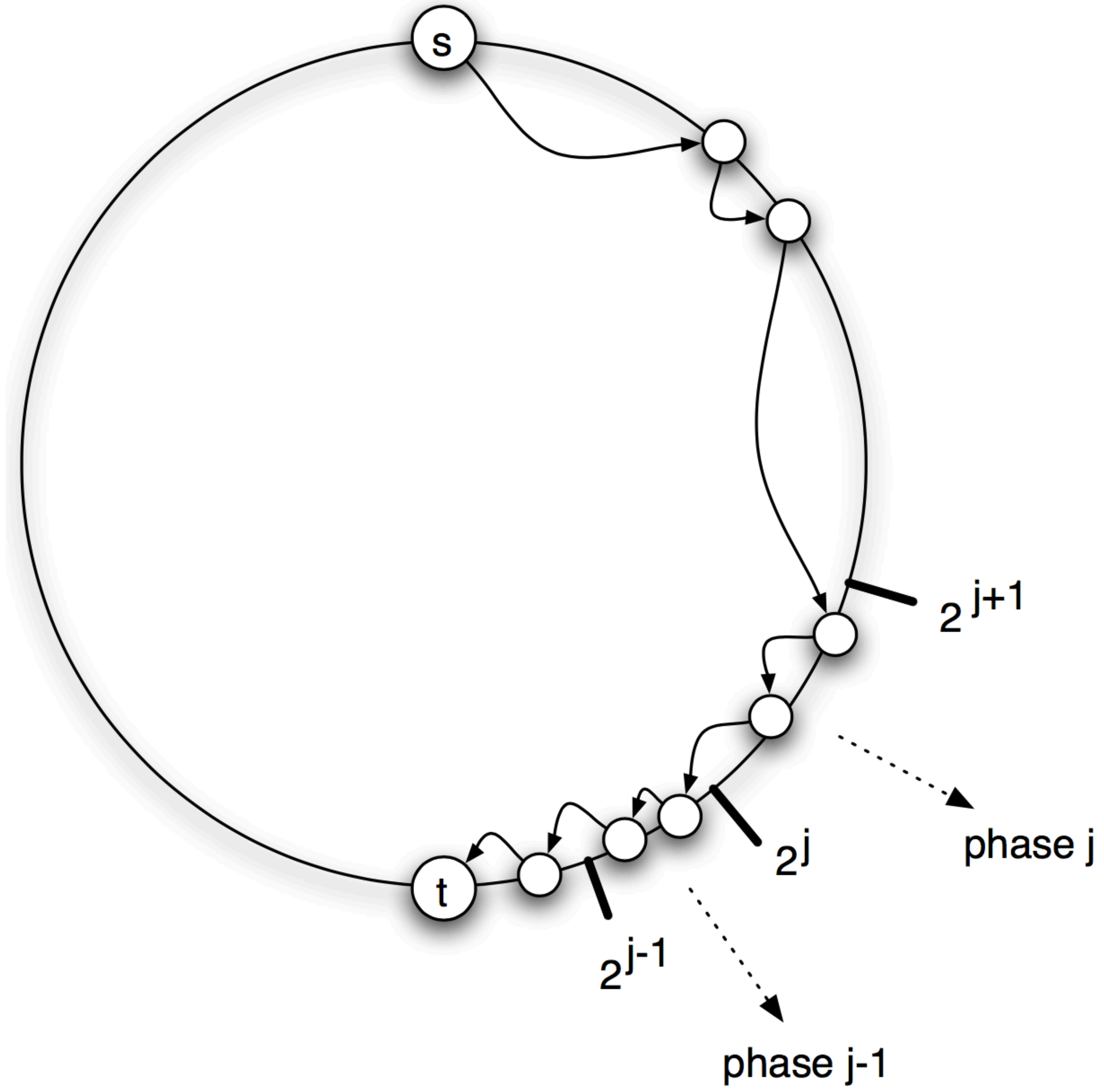
<u>Observation:</u> Notice that myopic search will always get closer to the target — even if your random link isn’t closer, one of your neighbours will be

We can split the search up into exponentially decreasing phases



Kleinberg's Model in 1-Dimension

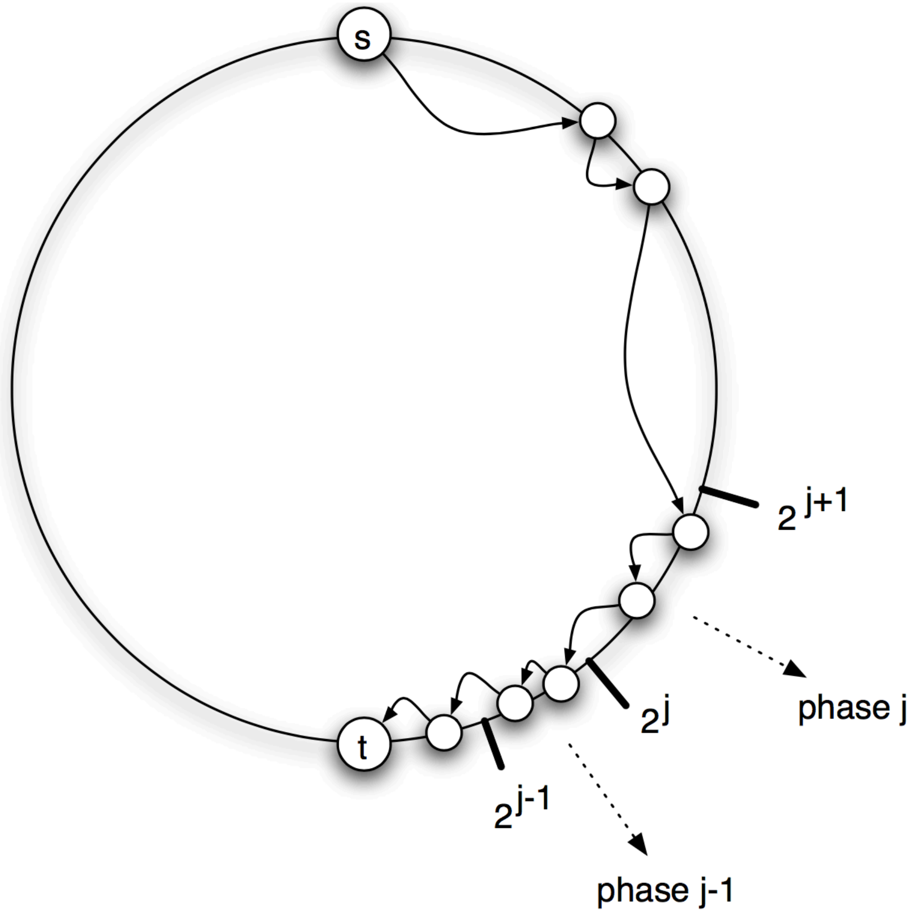
<u>Define</u>: Say the search is in *phase j* if the remaining distance to t is between 2^j and 2^{j+1}



Kleinberg's Model in 1-Dimension

<u>Define</u>: Say the search is in *phase j* if the remaining distance to t is between 2^j and 2^{j+1}

Distance	2^{j+3}	2^{j+2}	2^{j+1}	8	4	2	1	
Source								Target
Phase	j+2	j+1	j	2	1	0		



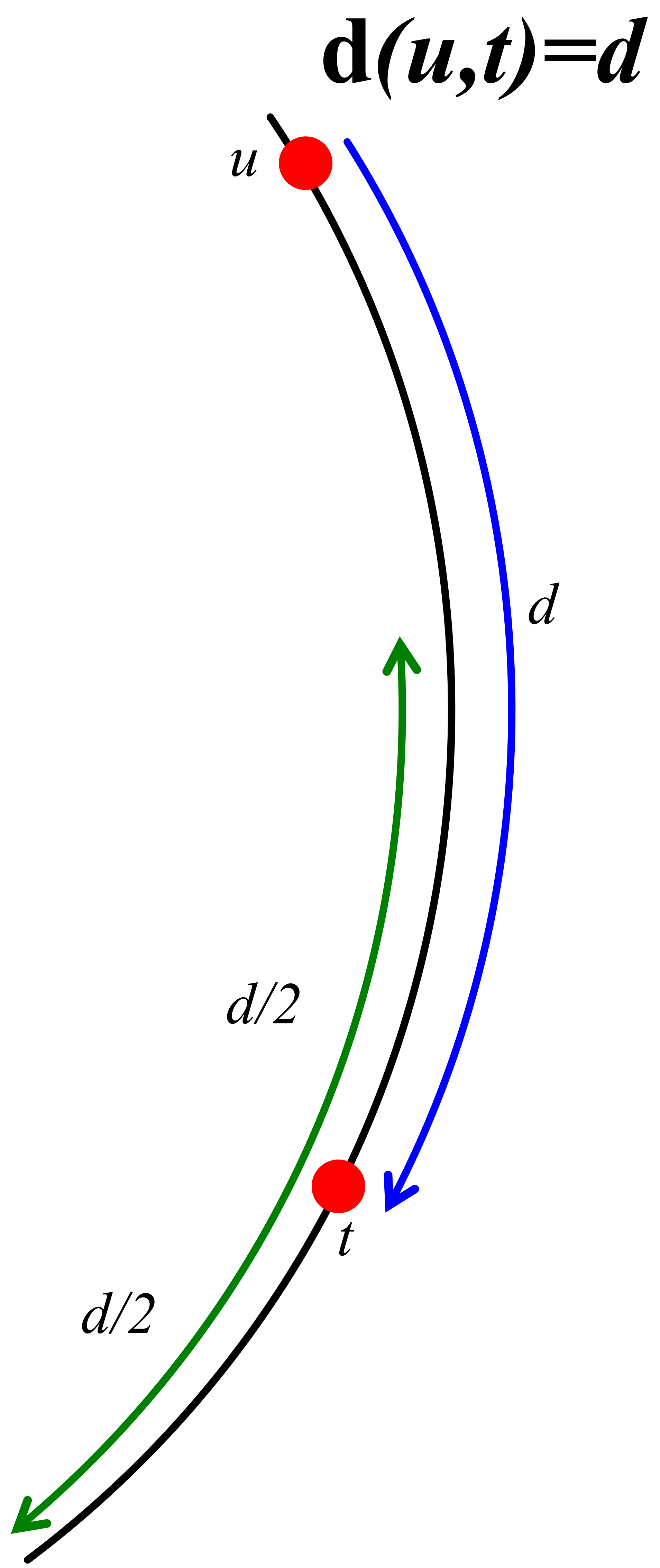
Kleinberg's Model in 1-Dimension

Call the remaining distance d(u, *t*) = *d*
Now consider the interval *I* of length *d* centered on the target

We will show that:

$$P\begin{pmatrix}\text{Long range}\\ \text{link from } u\\ \text{points to a}\\ \text{node in } I\end{pmatrix} = O\left(\frac{1}{\ln n}\right)$$

And remember
$\ln n = \log_e n$



Why is this nice? As *d* gets bigger,
I gets wider, but the prob. is independent of *d.*

Kleinberg's Model in 1-D

We need to calculate:

$$P(u \to v) = \frac{d(u,v)^{-1}}{\sum_{w \neq u} d(u,w)^{-1}}$$

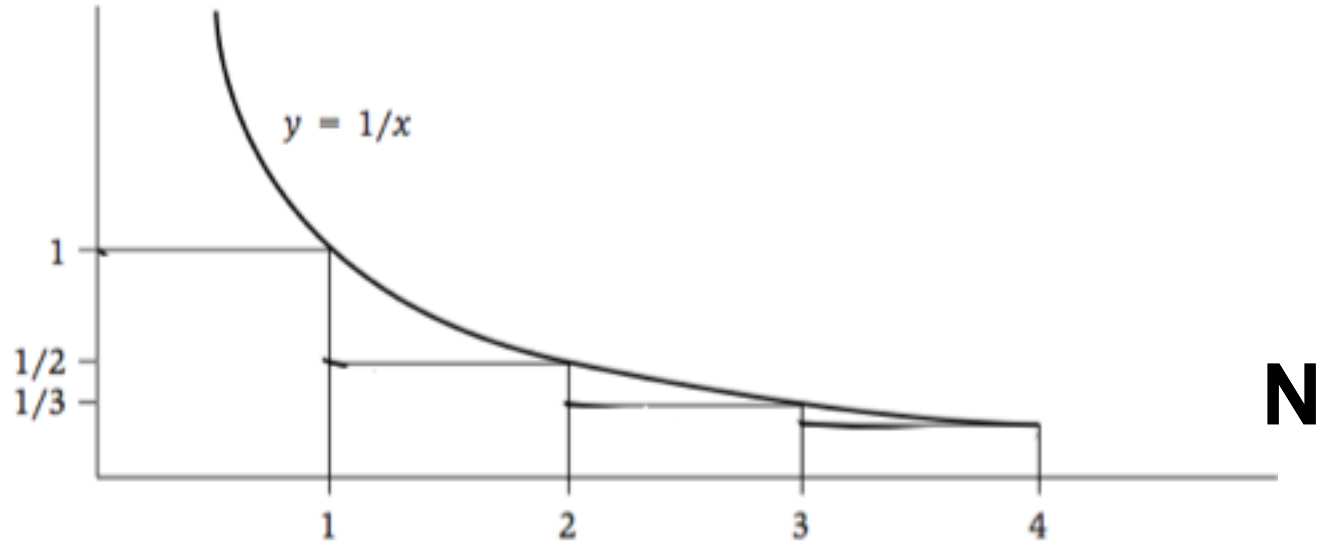
First: what is the normalizing constant?

$$\sum_{w \neq u} d(u,w)^{-1} = \sum 2\frac{1}{d} = 2\sum_{d=1}^{n/2}\frac{1}{d} \leq 2\ln n$$

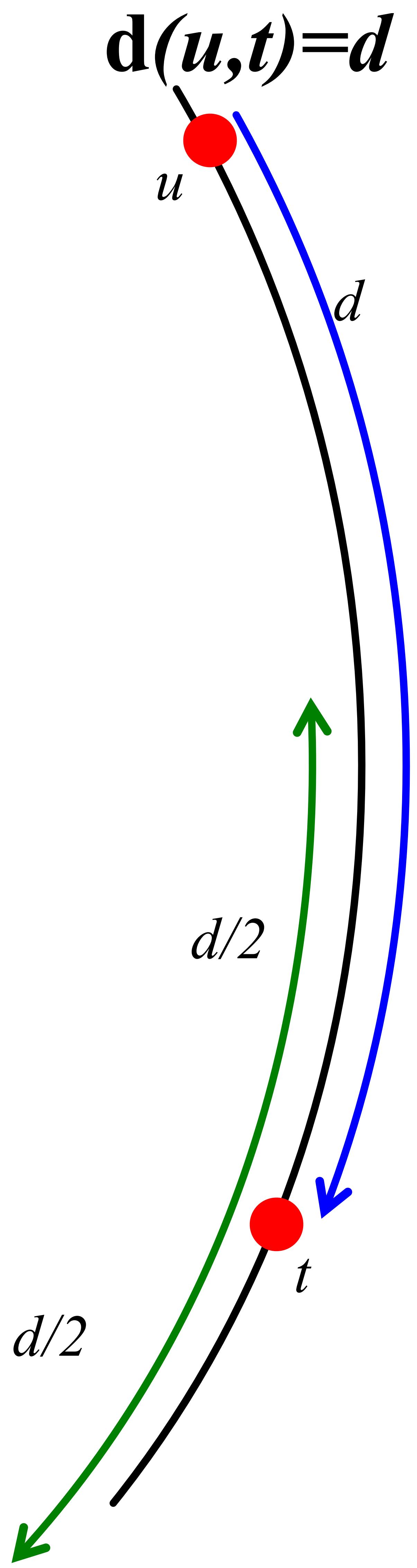
Every distance d from 1 to n/2

(At every distance d there are 2 nodes. Prob. of linking to one is 1/d, by definition)

(By the identity given below)



Note: $\sum_{d=1}^{n/2}\frac{1}{d} \leq 1 + \int_1^{n/2}\frac{dx}{x} = 1 + \ln\left(\frac{n}{2}\right) = \ln n$



Kleinberg's Model in 1-D

Now that we have the normalizing constant, we can explicitly calculate probability u's random long-range link points inside *I*:

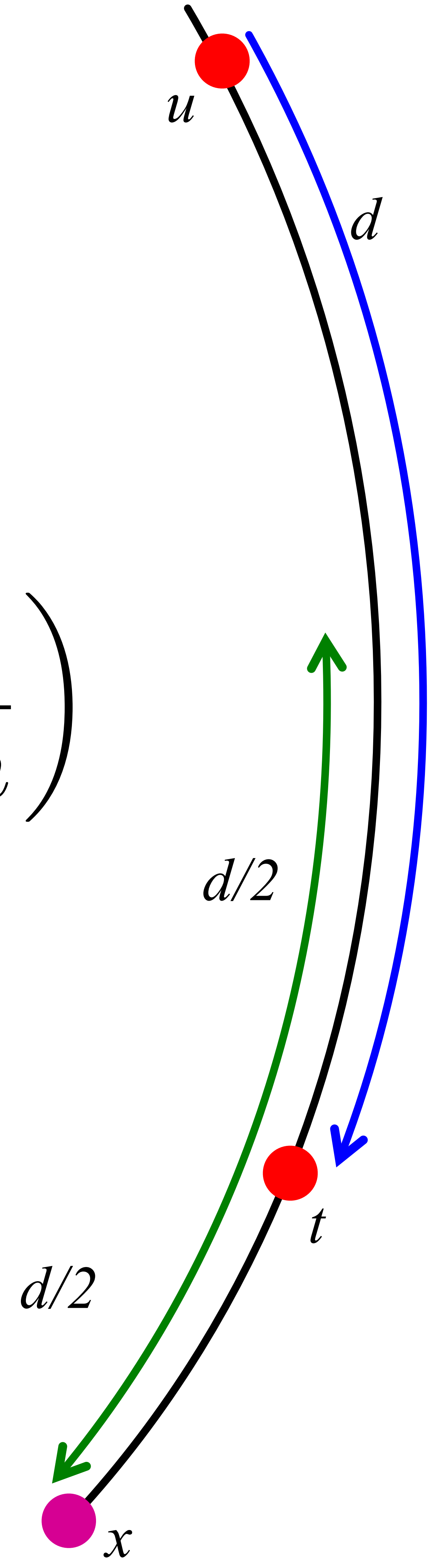
$$P(u \text{ points to } I) = \sum_{v \in I} P(u \to v) \geq \sum_{v \in I} \frac{d(u,v)^{-1}}{2 \ln n}$$

$$= \frac{1}{2 \ln n} \sum_{v \in I} \frac{1}{d(u,v)} \geq \frac{1}{2 \ln n} d \frac{2}{3d} = \frac{1}{3 \ln n} \in O\left(\frac{1}{\ln n}\right)$$

The biggest d(u,v) can be is 3d/2
So all terms ≥ 2/(3d)

There are d nodes in *I*

…and they're all closer than 3d/2



Note:
d(u,x)=3d/2

Kleinberg's Model in 1-D

We have:

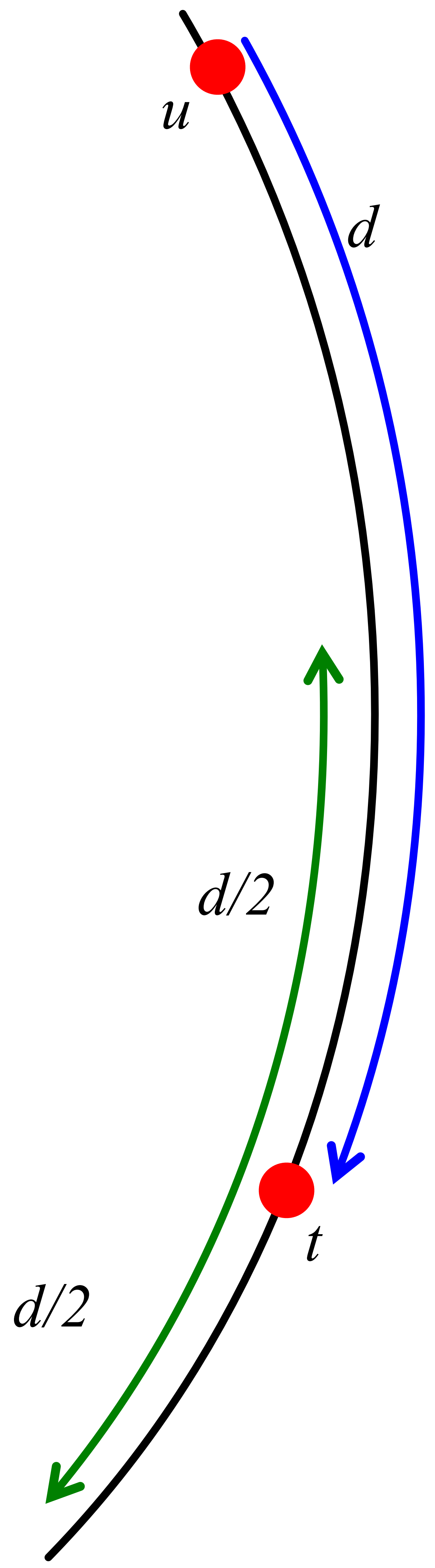
I ... interval of *d*/2 around ***t***

P(long link of *u* points to *I*) =1/ln(*n*)

In expected # of steps ≤ ln(*n*) you get into ***I***, and thus you halve the distance to ***t***

Distance can be halved at most $\log_2(n)$ times

So expected time to reach ***t***:
$O(\log_2(n)^2)$



Overview of the Results

Not searchable

Search time T:

$$O(n^{\alpha})$$

Watts-Strogatz

$$O(n^{\frac{2}{3}})$$

Erdős–Rényi

$$O(n)$$

Searchable

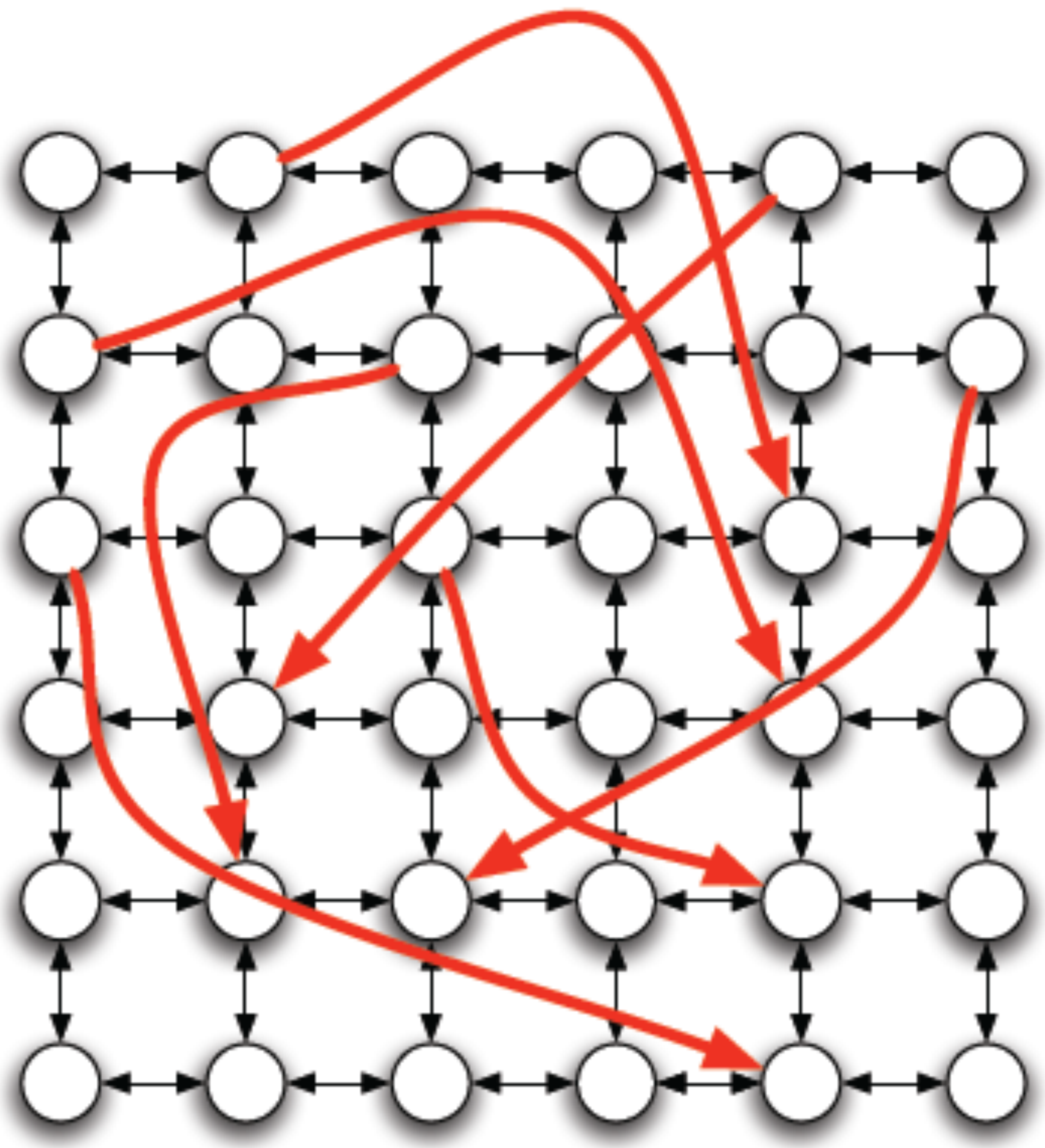
Search time T:

$$O((\log n)^{\beta})$$

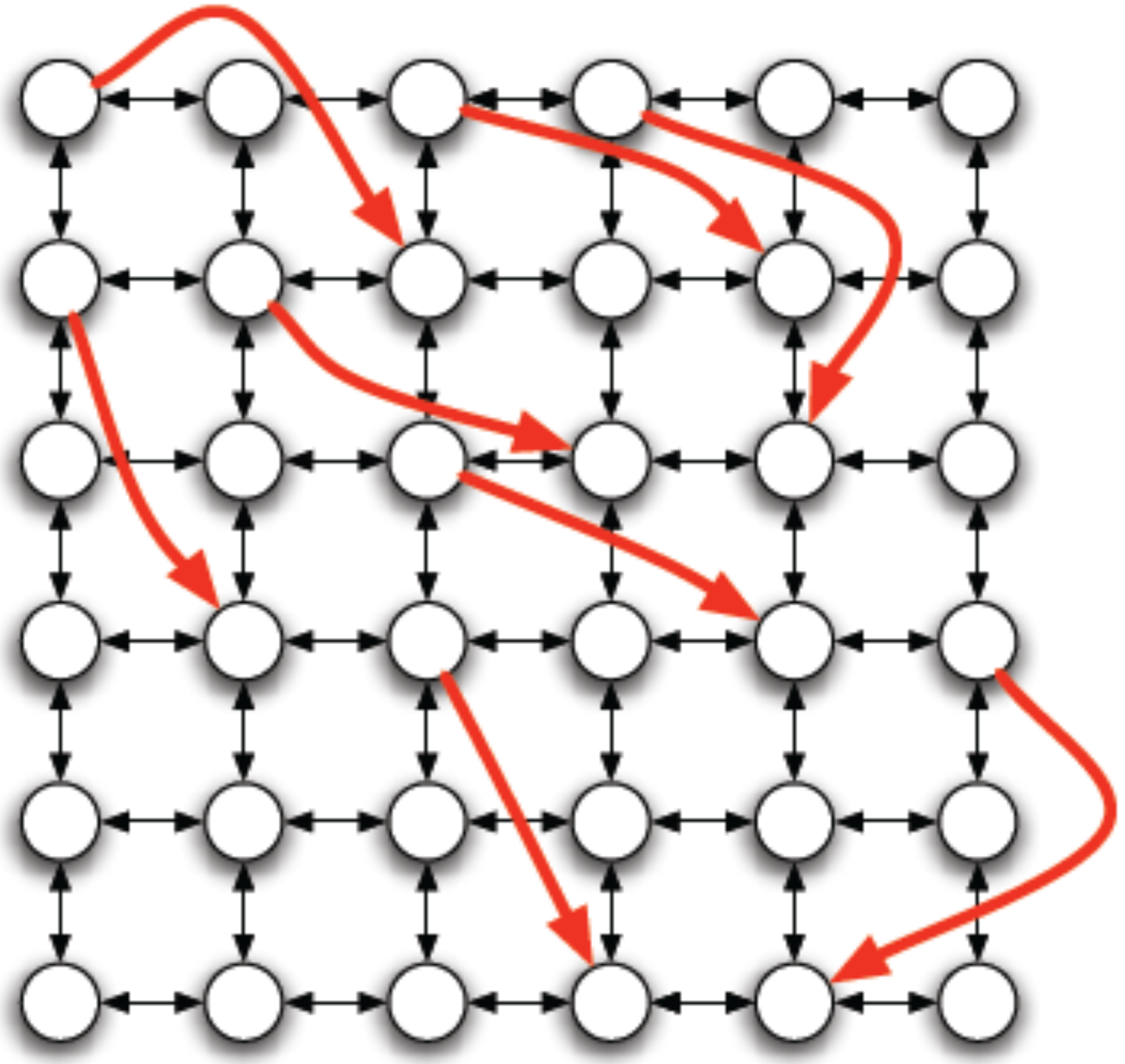
Kleinberg's model

$$O((\log n)^{2})$$

Intuition: Why Search Takes Long



Small α: too many long links



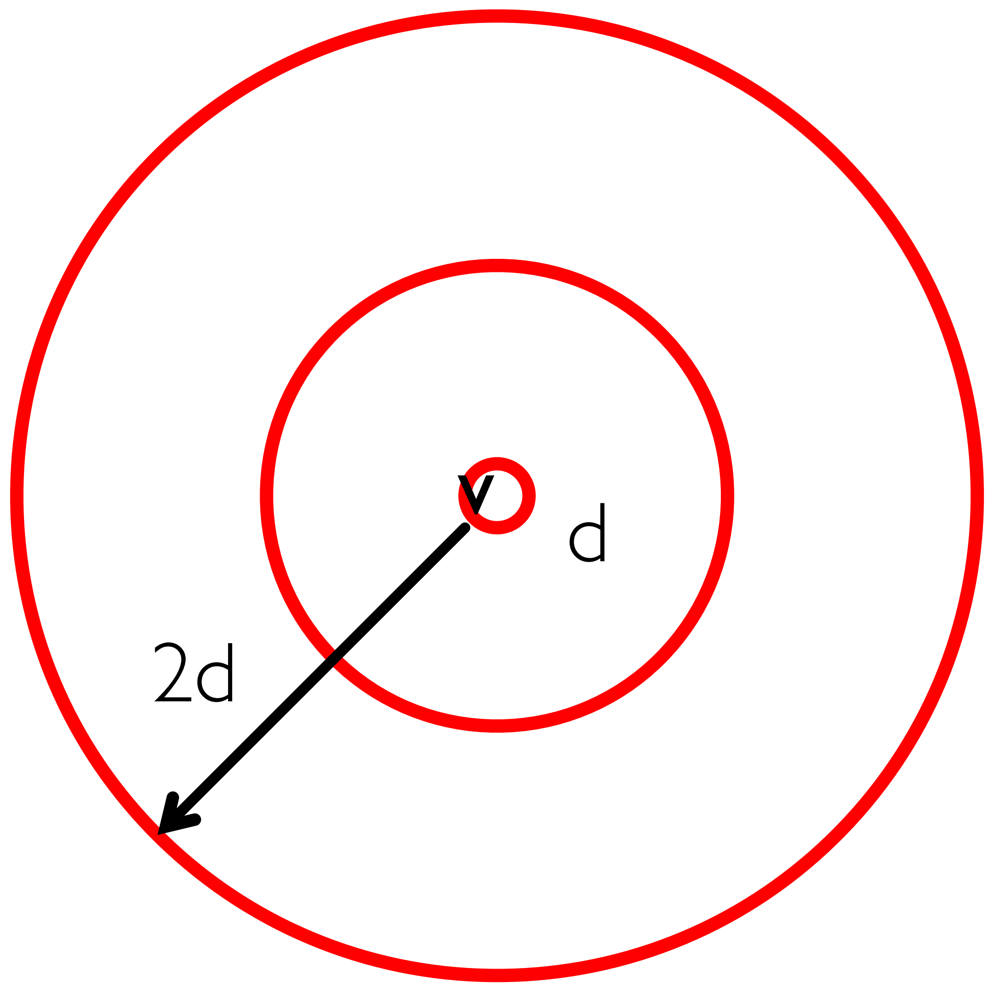
Big α: too many short links

Why Does It Work?

Why does $P(u \rightarrow v) \sim d(u,v)^{-dim}$ work?

Approx uniform over all "scales of resolution"

\# points at distance d grows as d^{dim}, prob. d^{-dim} of each edge
→ const. probability of a link, independent of d



Number of nodes is $\sim d^2$
Prob of linking to each one is $\sim d^{-2}$

Today: six degrees of separation

What are the basic properties properties of real social networks? **Short paths!**

How can we model them? **Path lengths**

Today:

- Milgram's experiment **Letters took 6 hops**
- Measuring path lengths in real-world networks
- Comparing with a baseline: G_{np} model ☺ ☹ ☹
- More realistic models: Watts&Strogatz model ☺ ☺ ☹
- More realistic models: Kleinberg's Decentralized search

Navigable!