The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

World Year of PHYSICS 2005

SMR.1656 - 11

School and Workshop on Structure and Function of Complex Networks

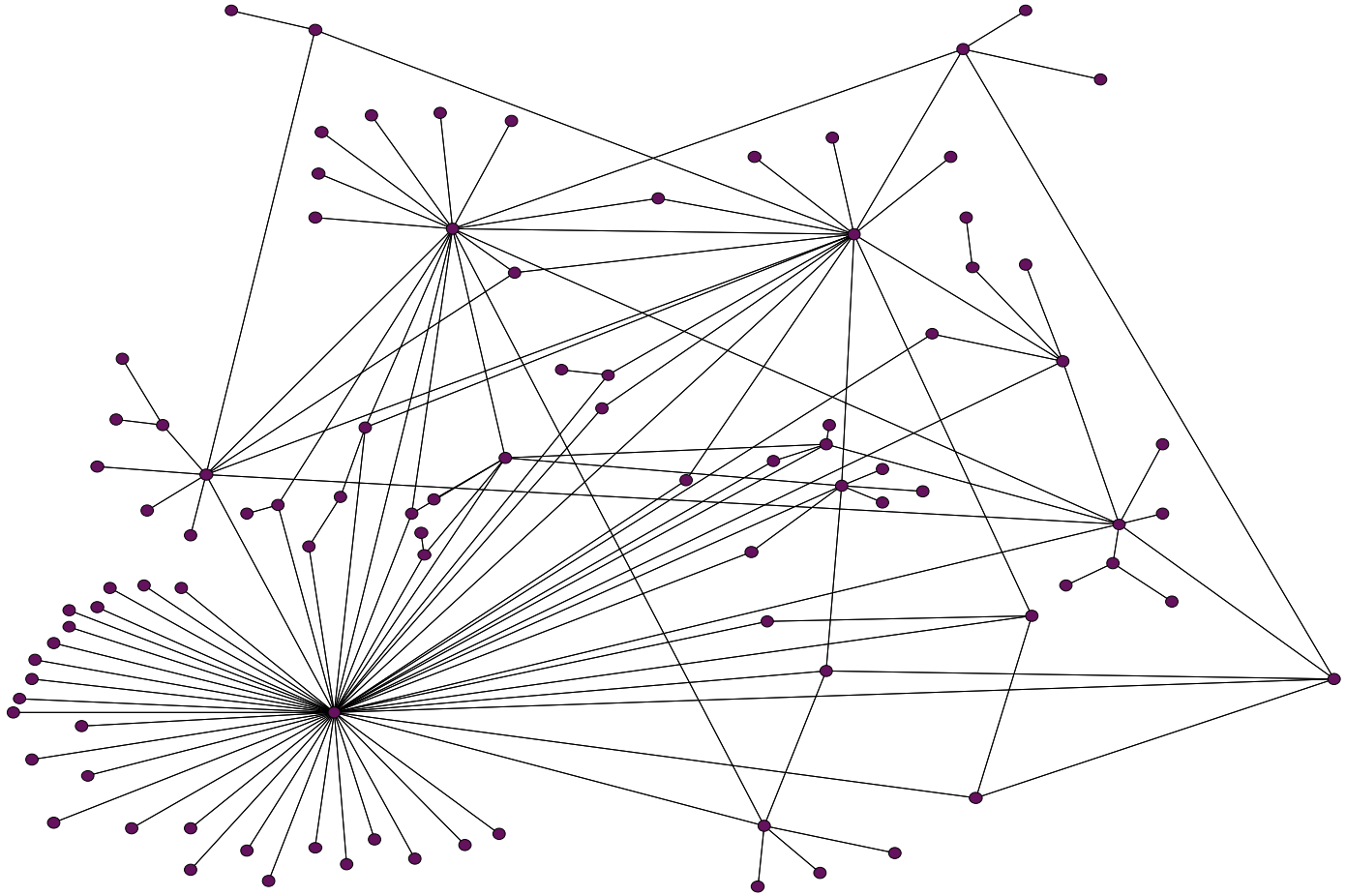
16 - 28 May 2005

Search in Random Networks

Lada A. ADAMIC
HP Laboratories
1501 Page Mill Road
MS 1139
Palo Alto CA 94024
U.S.A.

Search in Random Networks

Lada Adamic



School on the Structure and Function of Complex Networks, Trieste, 2005

Outline

Motivation

Power-law (PL) networks, social and P2P

Analysis of scaling of search strategies in PL networks

Simulation

artificial power-law topologies, real Gnutella networks

Comparison with existing P2P search strategies

Reflector, Morpheus

Path finding

Directed Search

Freenet

→ next lecture: Search in structured networks

How do we search?

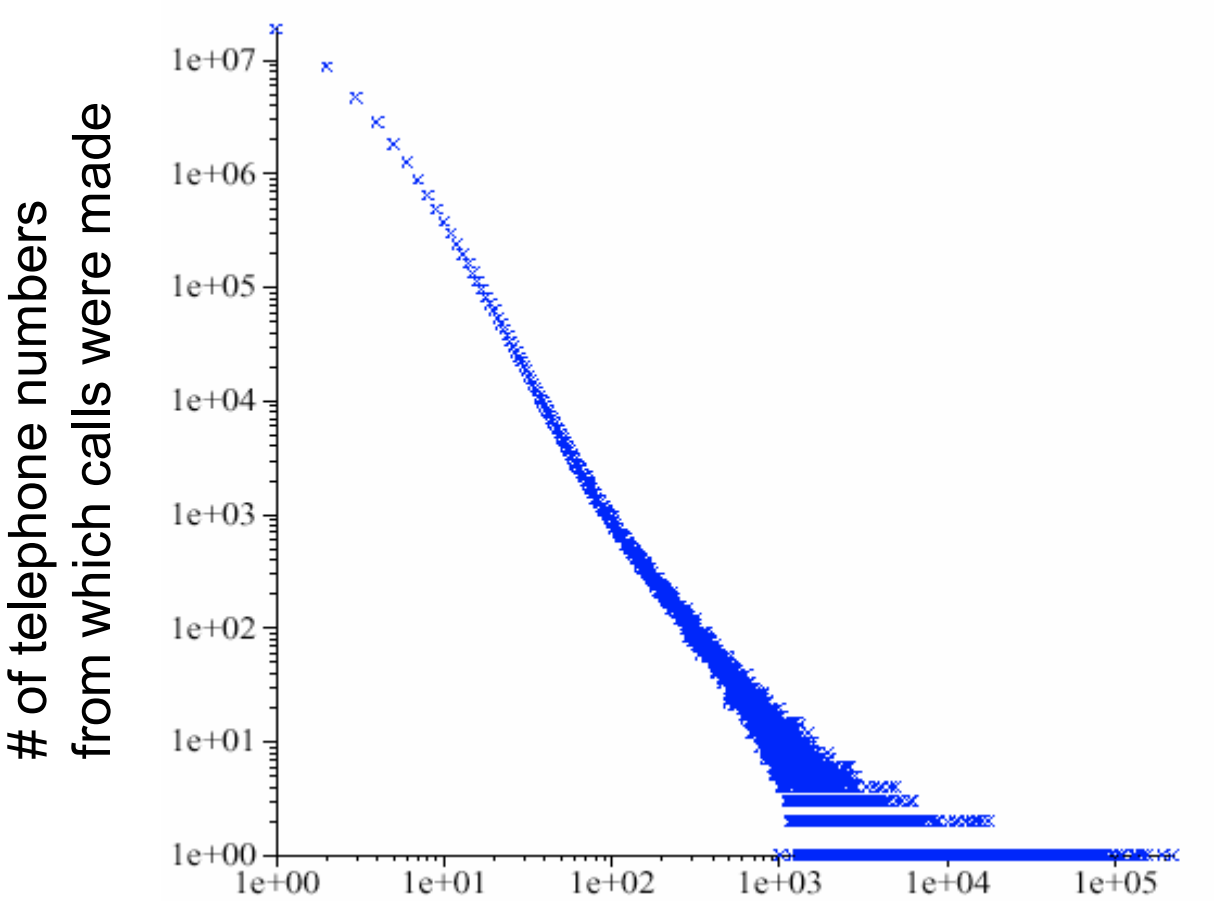
Mary

Who could introduce me to Richard Gere?

Bob

Jane

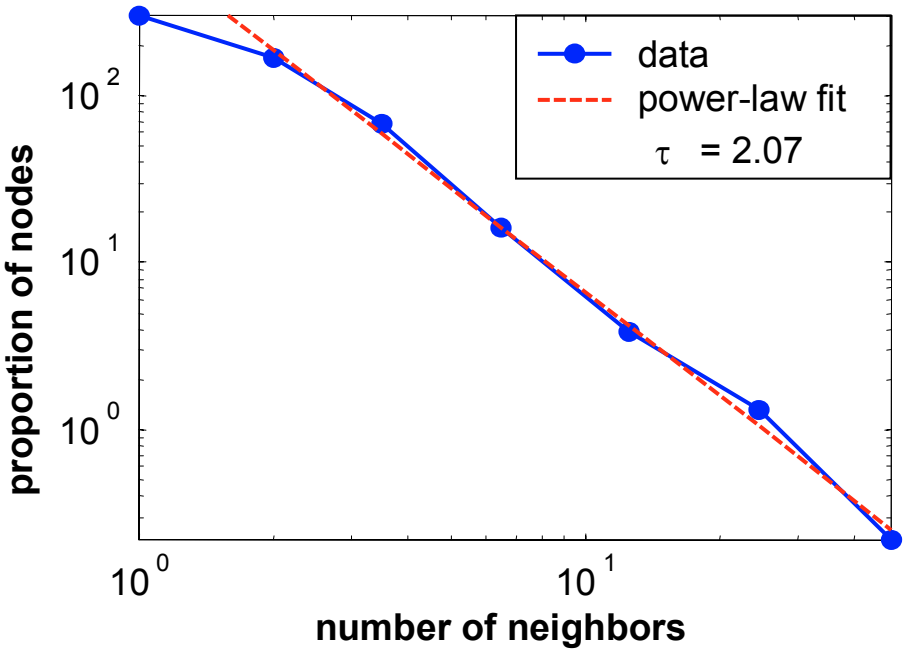
AT&T Call Graph



Aiello et al. STOC ‘00

Gnutella network

power-law link distribution



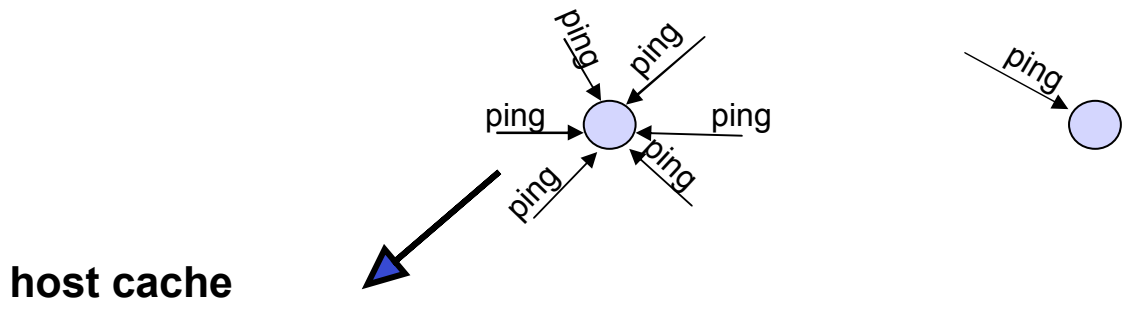
summer 2000,
data provided by Clip2

Preferential attachment model

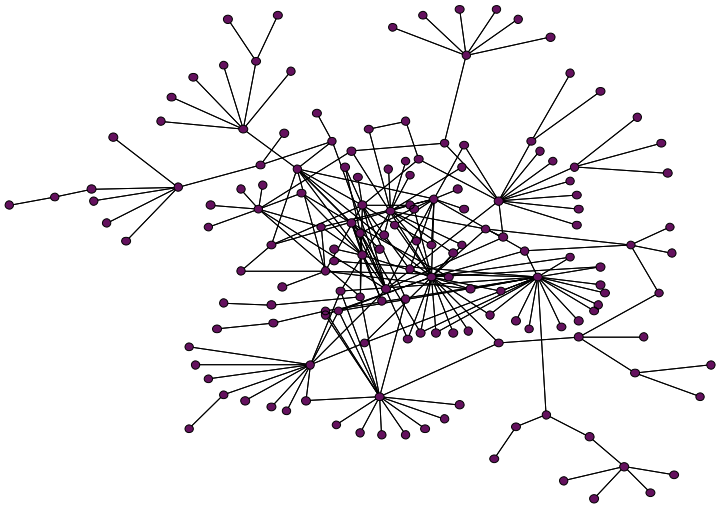
Nodes join at different times

The more connections a node has, the more likely it is to acquire new connections

Growth process produces power-law network



Gnutella and the bandwidth barrier

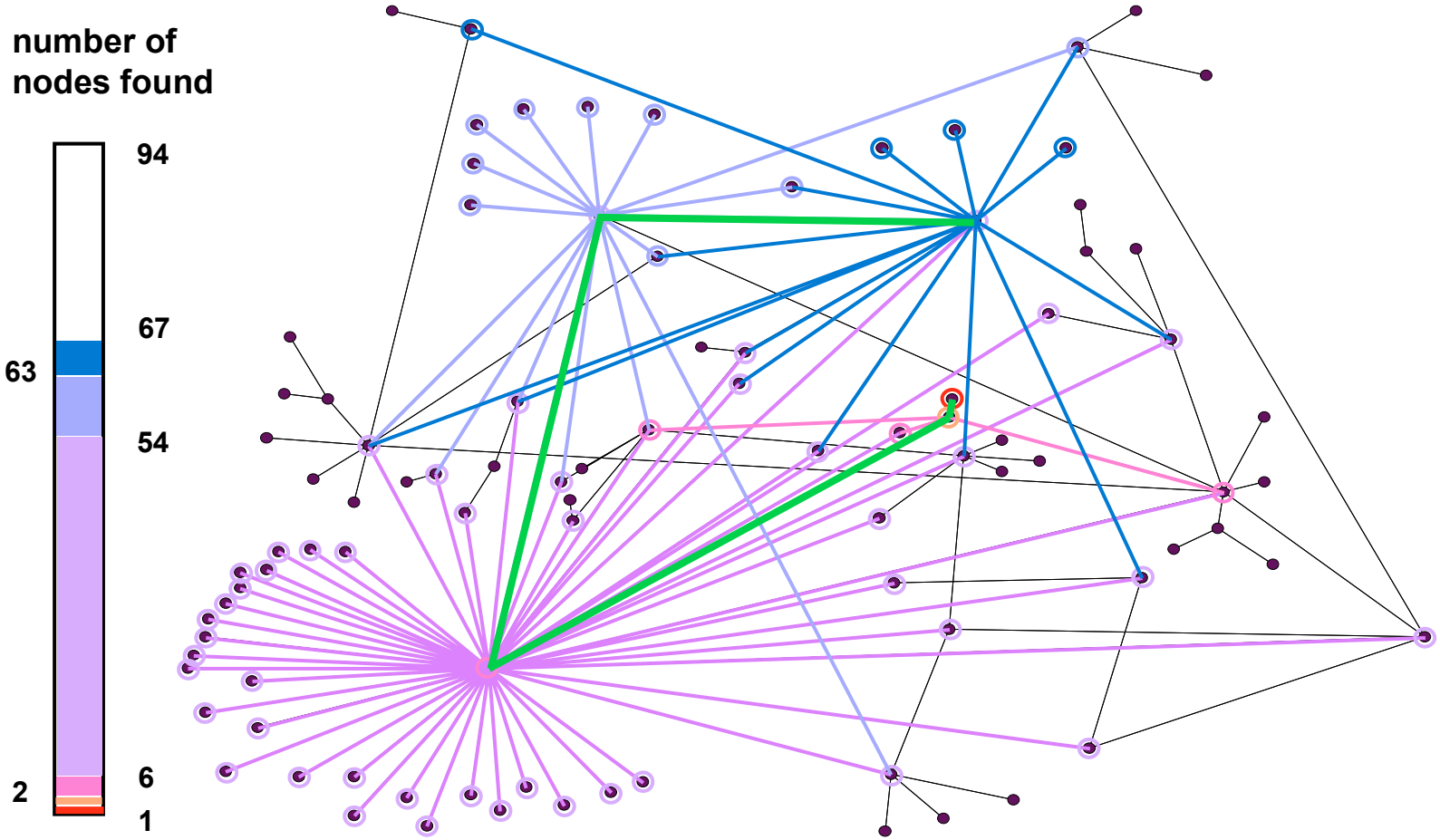


file sharing w/o a central index

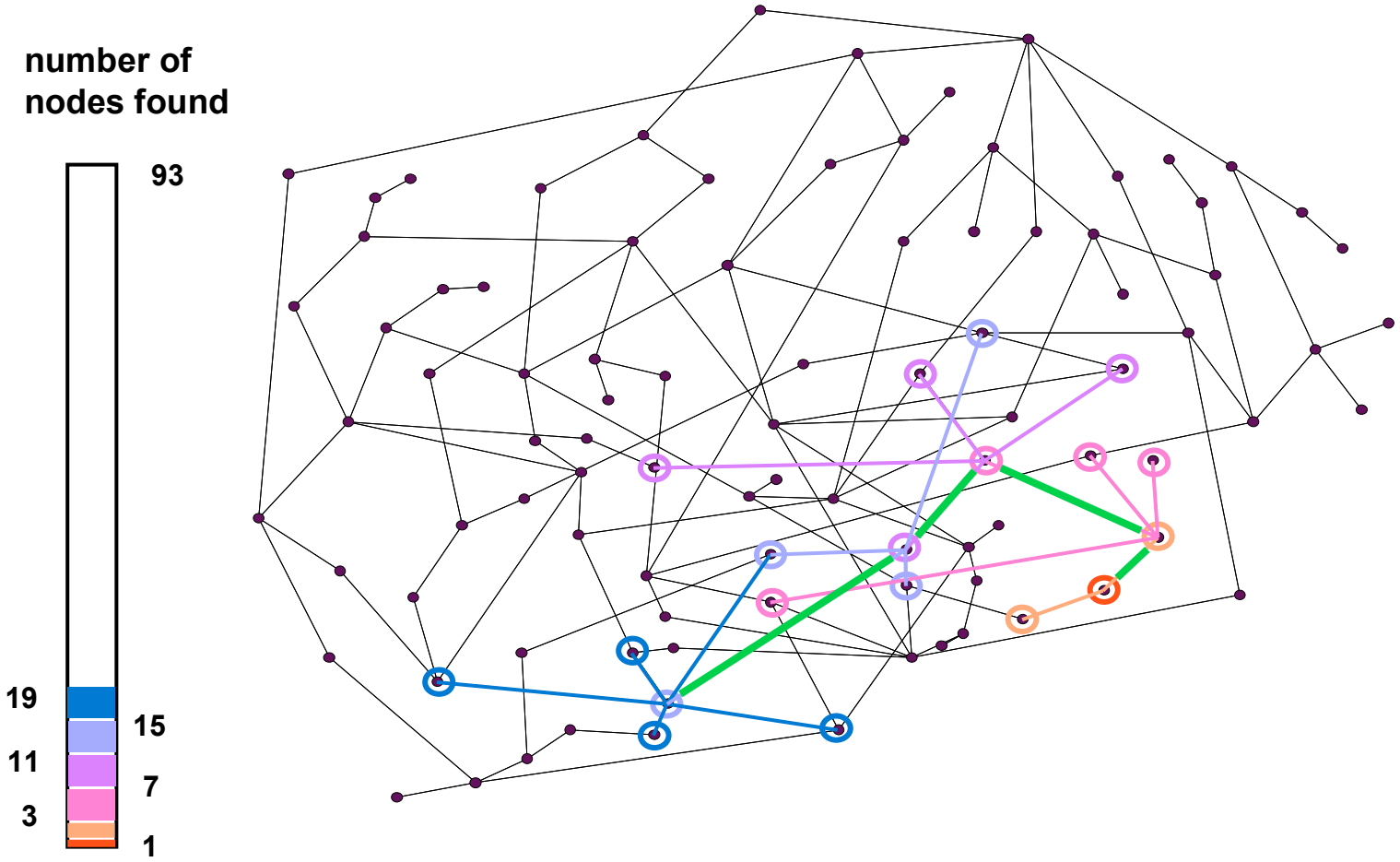
queries broadcast to every node within radius *ttl*
$\Rightarrow$ as network grows, encounter a bandwidth barrier
(dial up modems cannot keep up with query traffic,
fragmenting the network)

Clip 2 report
Gnutella: To the Bandwidth Barrier and Beyond
http://www.clip2.com/gnutella.html#q17

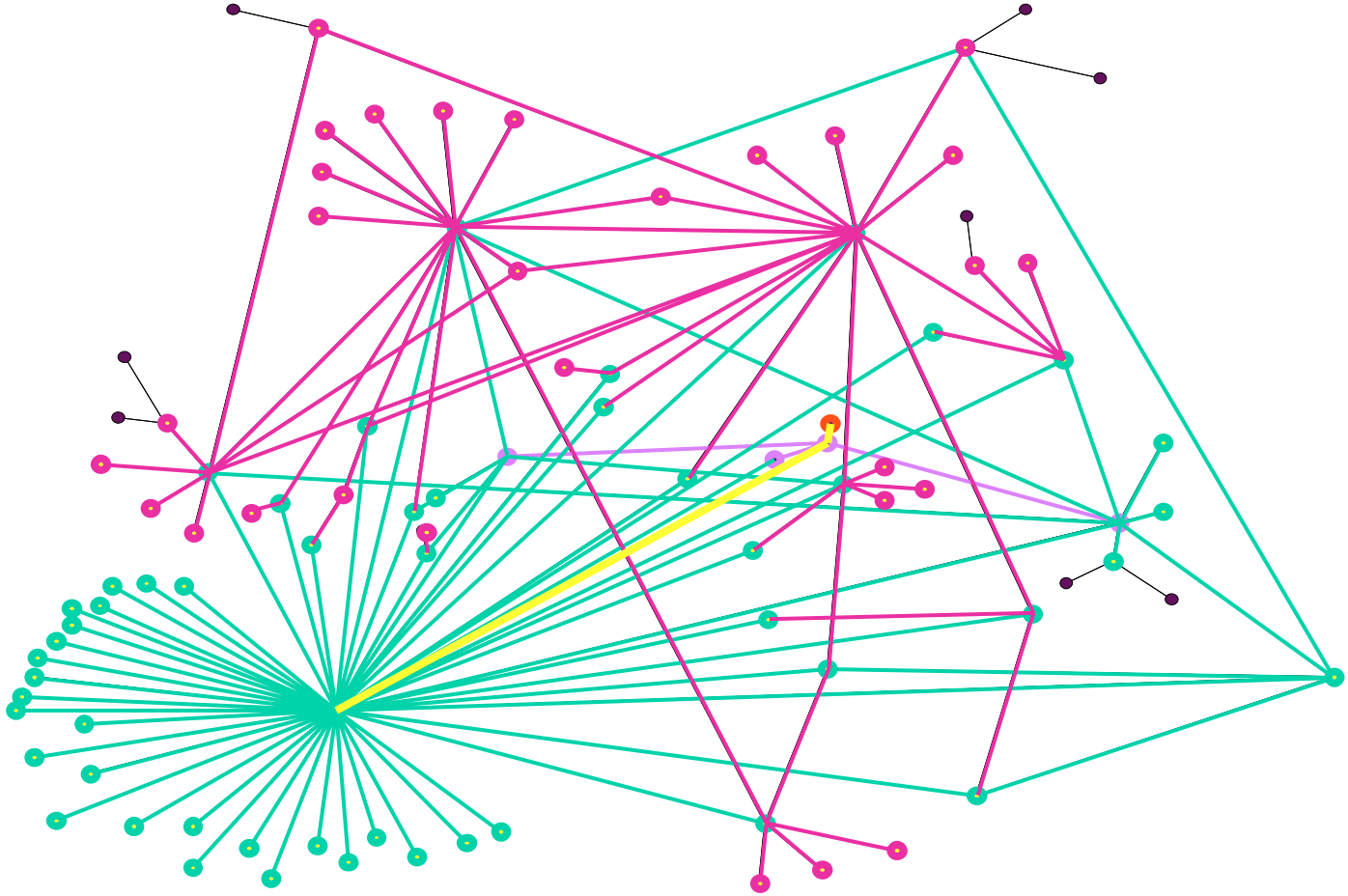
power-law graph



Poisson graph



Search with knowledge of 2nd neighbors



Outline of search strategy

pass query onto only **one** neighbor at each step

OPTIONS

requires that nodes sign query
- avoid passing message onto a node twice

requires knowledge of one's neighbors degree
- pass to the highest degree node

requires knowledge of one's neighbors neighbors
- route to 2nd degree neighbors

Generating functions

M.E.J. Newman, S.H. Strogatz, and D.J. Watts

'Random graphs with arbitrary degree distributions and their applications', PRE, cond-mat/0007235

Generating functions for degree distributions

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k$$

Useful for computing moments of degree distribution,

component sizes, and average pathlengths

Introducing cutoffs

$k_{\max} < N - 1$ a node cannot have more connections than there are other nodes

This is important for exponents close to 2

$$\sum_{1}^{\infty} p_k = \sum_{1}^{\infty} C_\tau \frac{1}{x^\tau} = 1 \qquad C_2 = \frac{6}{\pi^2}$$

$$p(k > 1000, \tau = 2) = \sum_{1000}^{\infty} p_k \sim 0.001$$

Probability that none of the nodes in a 1,000 node graph has 1000 or more neighbors:

$$(1 - p(k > 1000, \tau = 2))^{1000} \sim 0.36$$

without a cutoff, for $\tau = 2$
have > 50% chance of observing a node with more neighbors than there are nodes

for $\tau = 2.1$, have a 25% chance

Selecting from a variety of cutoffs

1. $k_{\max} < N$

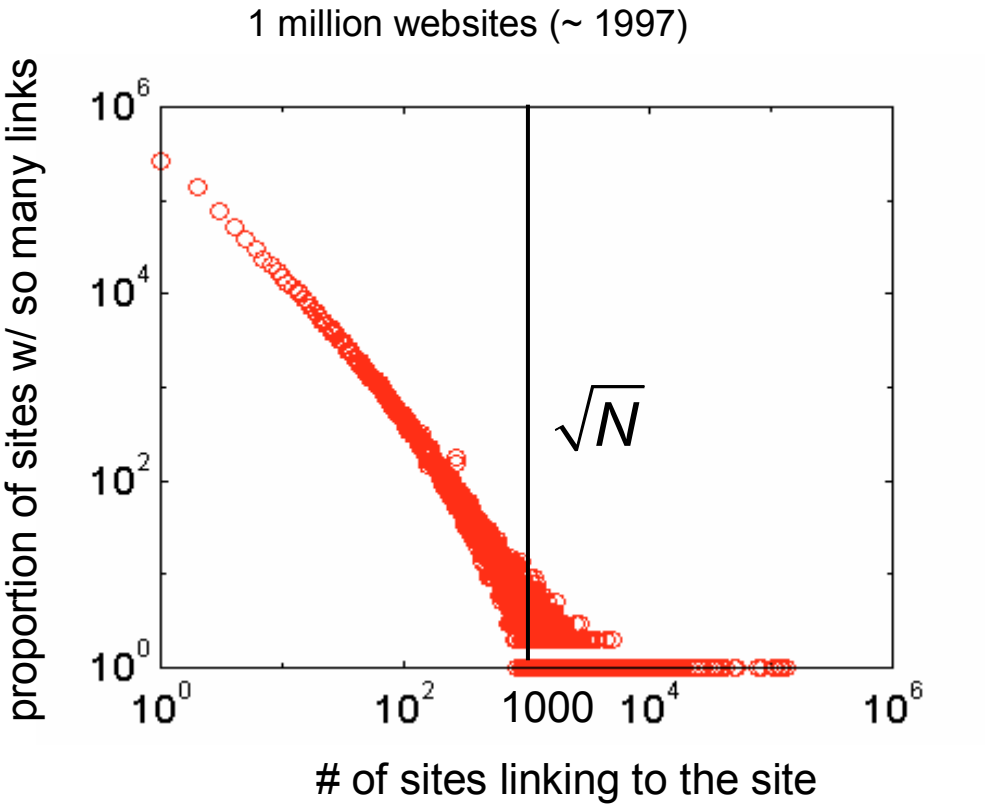
2. $p_k = Ck^{-\tau} e^{-k/\kappa}$ Newman et al.

3. $p_k = \begin{cases} Ck^{-\tau} & k < (CN)^{1/\tau} \\ 0 & \text{otherwise} \end{cases}$

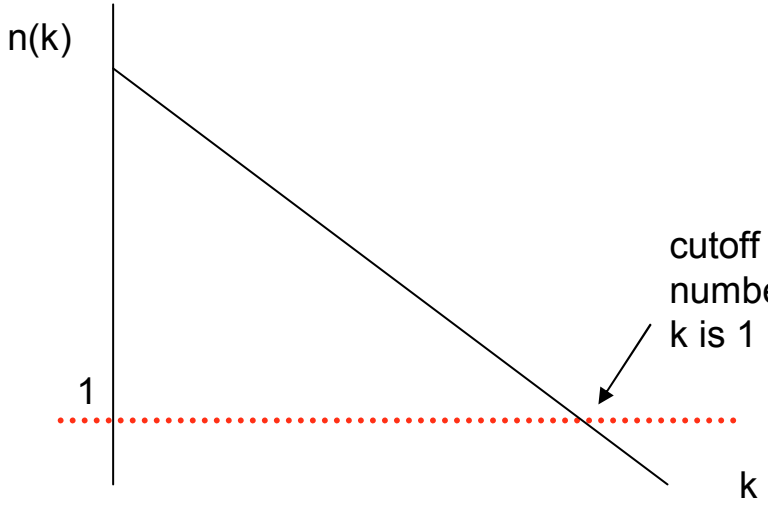
Aiello et al.

Generating Function

$$G_0(x) = C \sum_{k=1}^{(CN)^{1/\tau}} k^{-\tau} x^k$$



Aiello's 'conservative' vs. Havlin's 'natural' cutoff

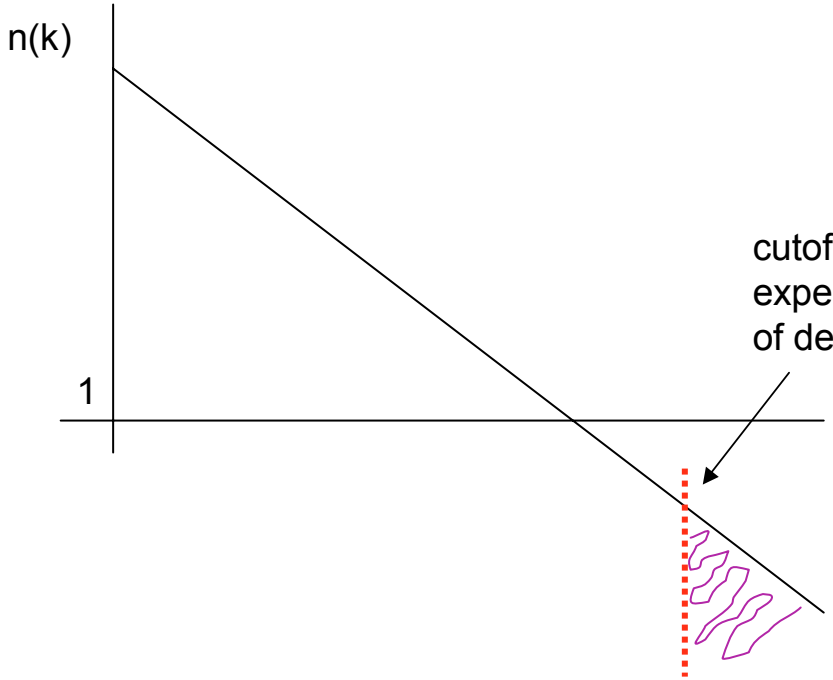


cutoff where expected number of nodes of degree k is 1

$$N * p_k = 1$$

$$Ck^{-\tau} = N^{-1}$$

$$k \sim N^{\frac{1}{\tau}}$$



cutoff so that expected number of nodes of degree > k is 1

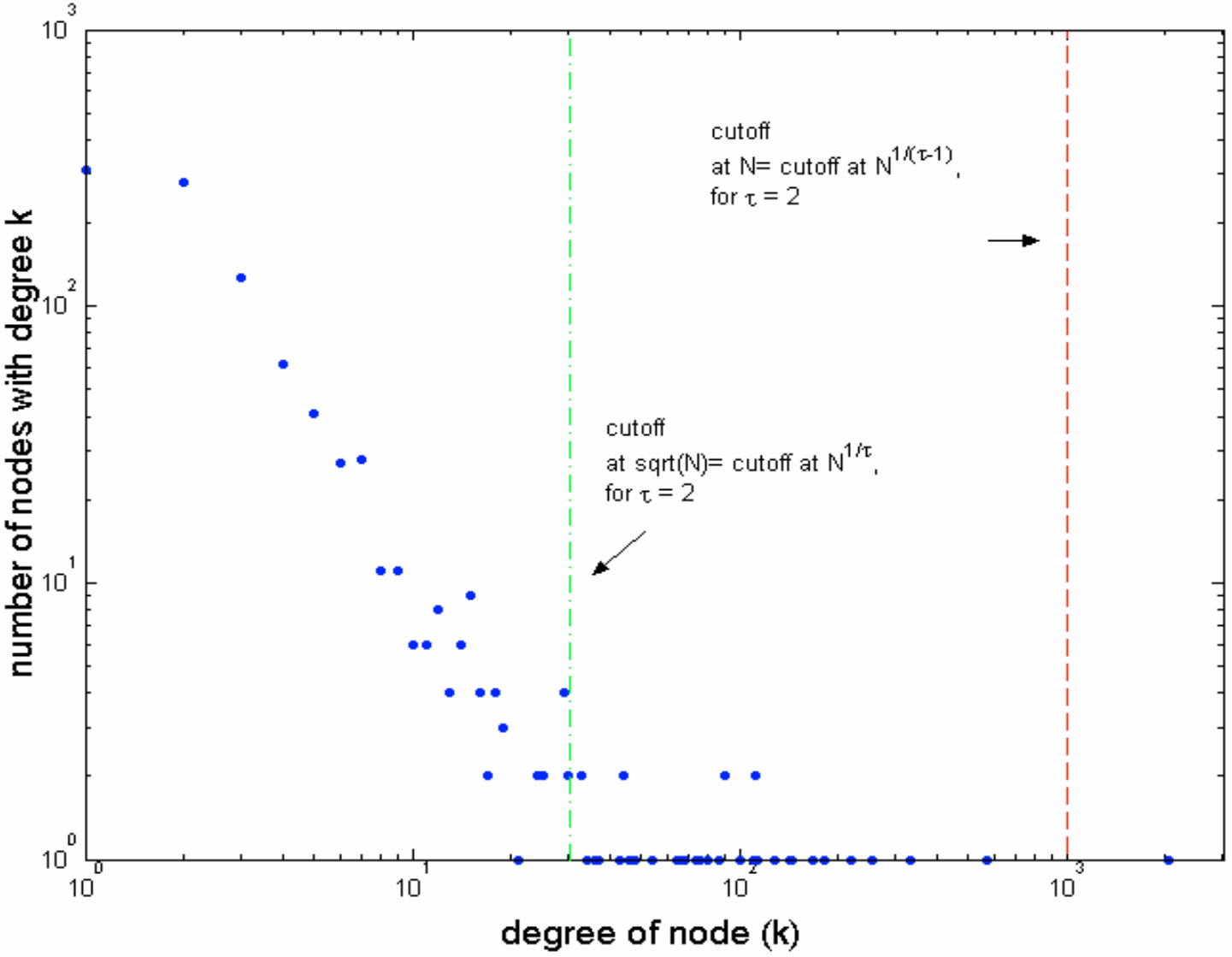
$$N * \sum_{k=k_{\max}}^{\infty} p_k = 1$$

$$\int_{k=k_{\max}}^{\infty} ck^{-\tau} \sim N^{-1}$$

$$k_{\max}^{1-\tau} \sim N^{-1}$$

$$k_{\max} \sim N^{\frac{1}{\tau-1}}$$

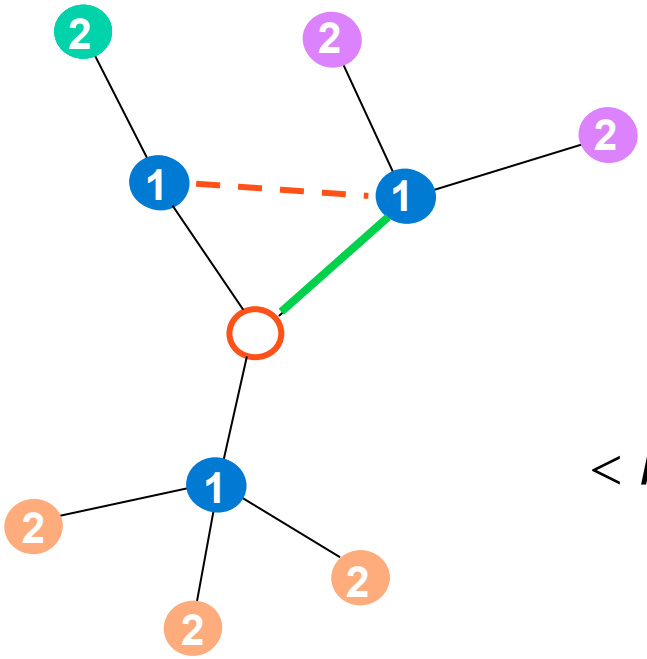
The imposed cutoff can have a dramatic effect on the properties of the graph



Generating functions for degree distributions

Random graphs with arbitrary degree distributions and their applications

by Newman, Strogatz & Watts



$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k$$ is a generating function

$$p_k \sim k^{-\tau}$$ is the probability that a randomly chosen vertex has degree k

$$< k >= \sum_k k p_k = G_0'(1)$$ is the expected degree of a randomly chosen vertex

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)}$$ is the distribution of remaining outgoing edges following and edge

$$z_2 = G_0'(1) G_1'(1)$$ is the expected number of second degree neighbors

assuming neighbors don't share edges

search with knowledge of first neighbors

$$G_0(x) = c\sum_{1}^{k_{max}} k^{-\tau} x^k$$

Generating function with cutoff

$$G_0'(x) = \frac{\partial}{\partial x} G_0(x) = c\sum_{1}^{k_{max}} k^{1-\tau} x^{k-1}$$

Average degree of vertex

$$G_0'(1) = < k > = c\sum_{1}^{k_{max}} k^{1-\tau} \approx \int_{1}^{k_{max}} k^{1-\tau} dk = \frac{1}{\tau - 2}\left(1 - k_{max}^{2-\tau}\right)$$

$$G_1'(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{c}{G_0'(1)} \frac{\partial}{\partial x} \sum_{1}^{k_{max}} k^{1-\tau} x^{k-1}$$

Average number of neighbors following an edge

$$= \frac{c}{G_0'(1)} \sum_{2}^{k_{max}} k^{1-\tau} (k-1) x^{k-2}$$

$$G_1'(1) = \frac{1}{G_0'(1)} \frac{k_{max}^{3-\tau}(\tau - 2) - 2^{2-\tau}(\tau - 1) + k_{max}^{2-\tau}(3 - \tau)}{(\tau - 2)(3 - \tau)}$$

constant in N

for $2<\tau<3$, and $k_{max} \sim N^a$, decreases with N

search with knowledge of first neighbors (cont'd)

$$z_{1B} = G_1^{'}(1) \approx \frac{1}{G_0^{'}(1)} \frac{k_{\max}^{3-\tau}}{(3-\tau)} = \frac{\tau-2}{1-k_{\max}^{2-\tau}} \frac{k_{\max}^{3-\tau}}{(3-\tau)} \approx k_{\max}^{3-\tau}$$

In the limit t->2, $$G_1^{'}(1) \approx \frac{k_{\max}}{\log(k_{\max})}$$

Let's for the moment ignore the fact that as we do a random walk, we encounter neighbors that we've seen before

s = number of steps = $$\frac{N}{z_{1B}}$$

Search time with different cutoffs

If $k_{max} = N$,

$$s(\tau) \sim \frac{N}{k_{max}^{3-\tau}} = \frac{N}{N^{3-\tau}} = N^{\tau-2}, 2 < \tau < 3$$

$$s(2.1) \sim N^{0.1}$$

$$s \sim \frac{N \log(k_{max})}{k_{max}} = \log(N), \tau = 2$$

If $k_{max} = N^{1/(\tau-1)}$,

$$s(\tau) \sim \frac{N}{k_{max}^{3-\tau}} = \frac{N}{N^{\frac{3-\tau}{\tau-1}}} = N^{2\frac{\tau-2}{\tau-1}}, 2 < \tau < 3$$

$$s(2.1) \sim N^{0.18}$$

$$s(2) \sim \frac{N \log(k_{max})}{k_{max}} = \log(N)$$

search with knowledge of first neighbors (cont'd)

If $k_{max} = N^{1/\tau}$,

$$s \sim \frac{N}{k_{\max}^{3-\tau}} = \frac{N}{(N^{\frac{1}{\tau}})^{3-\tau}} = N^{2-3/\tau}, 2 < \tau < 3$$

So the best we can do is $\sqrt{N}$ for exponents close to 2

2nd neighbor random walk, ignoring overlap:

$$n_s = z_{2B}(N)$$

$$S \sim \frac{N}{z_{2B}(N)}$$

$$z_{2B} = \left[\frac{\partial}{\partial x} G_1(G_1(x))\right]_{x=1} = \left[G_1'(1)\right]^2 = \left[\frac{\tau-2}{1-k_{\max}^{2-\tau}} \frac{k_{\max}^{3-\tau}}{(3-\tau)}\right]^2$$

$$S(N,\tau) \sim N^{3(1-2/\tau)} \qquad S(N,\tau=2.1) \sim N^{0.15}$$

Following the degree sequence

Go to highest degree node, then next highest, … etc.

$$z_{1D} = \int_{k_{\max}-a}^{k_{\max}} Nk^{1-\tau} dk \sim Nak_{\max}^{1-\tau}$$

a ~ s = # of steps taken

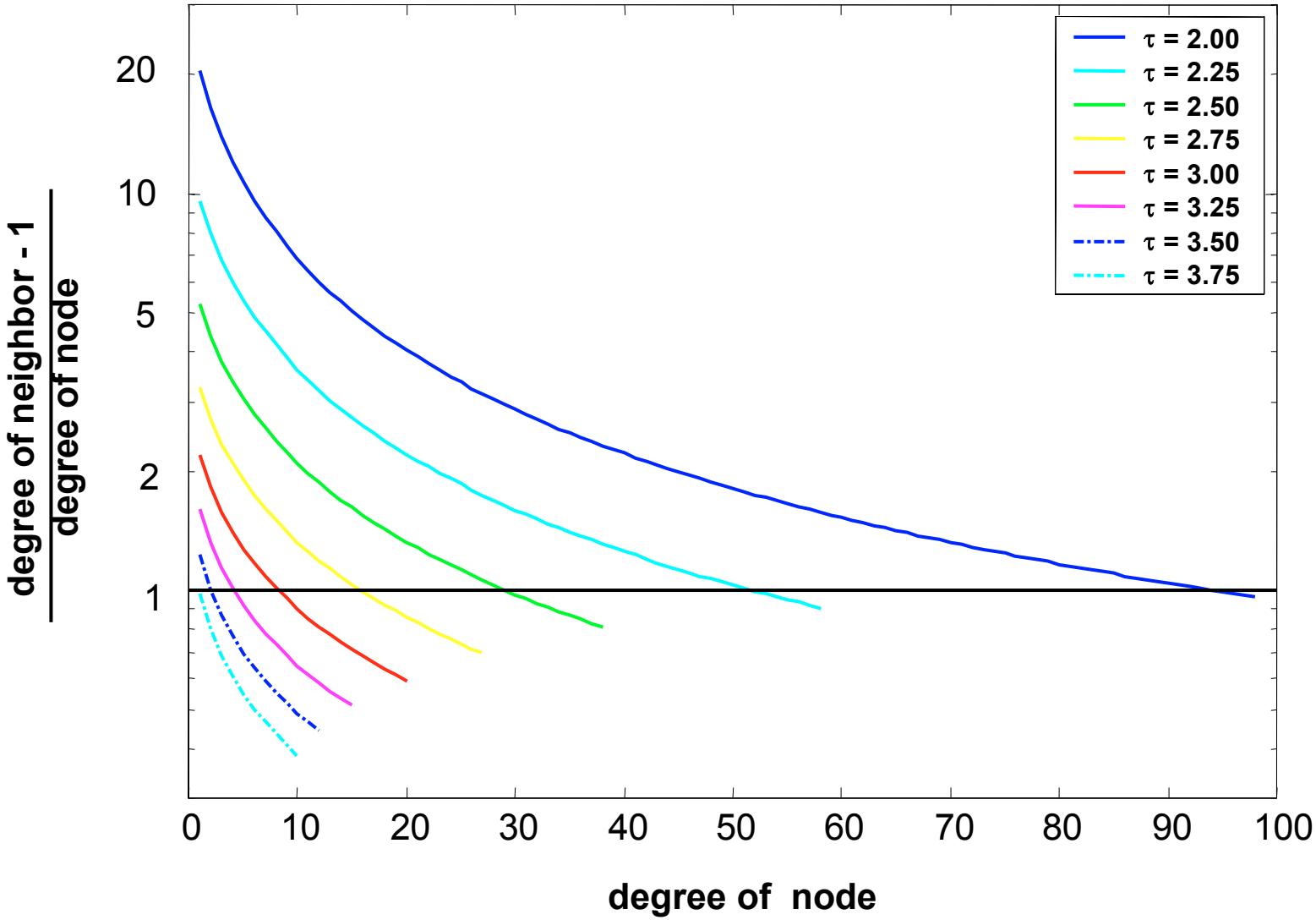
2nd neighbors, ignoring overlap:

$$z_{1D} G_1^{'}(x) \sim Nak_{\max}^{2(2-\tau)}$$

$$s \sim k_{\max}^{2(\tau-2)} \sim N^{2-4/\tau}$$

$$S_{\deg}\left(N, \tau = 2.1\right) = N^{0.1}$$

Ratio of the degree of a node to the expected degree of its highest degree neighbor for 10,000 node power-law graphs of varying exponents



Exponents τ close to 2 required to search effectively

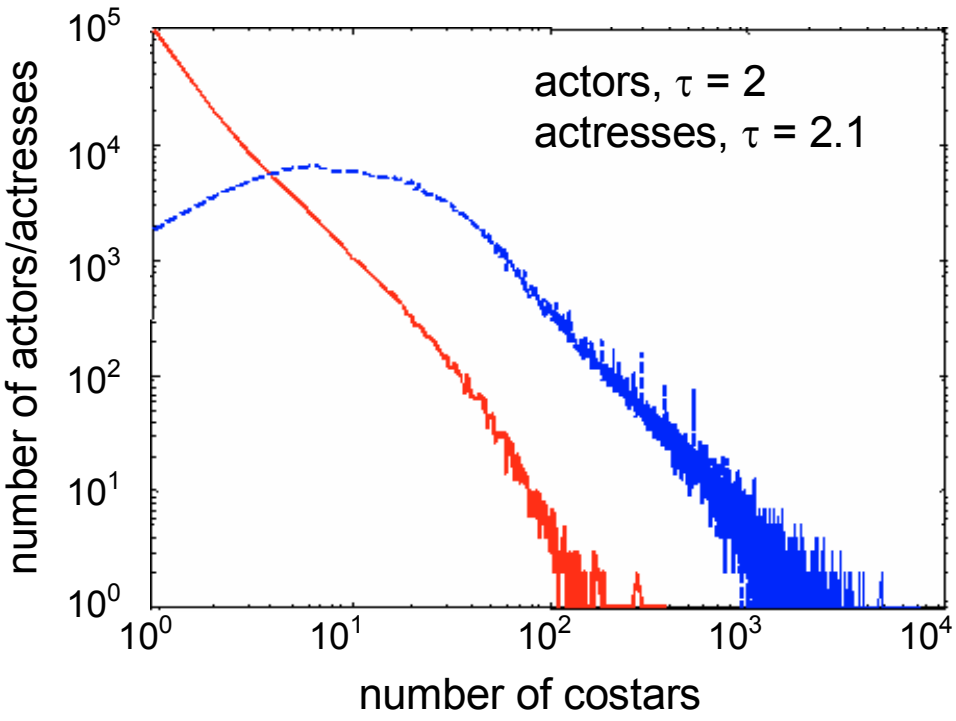
Gnutella

World Wide Web, $\tau \sim 2.0\text{-}2.3$,
high degree nodes: directories, search engines

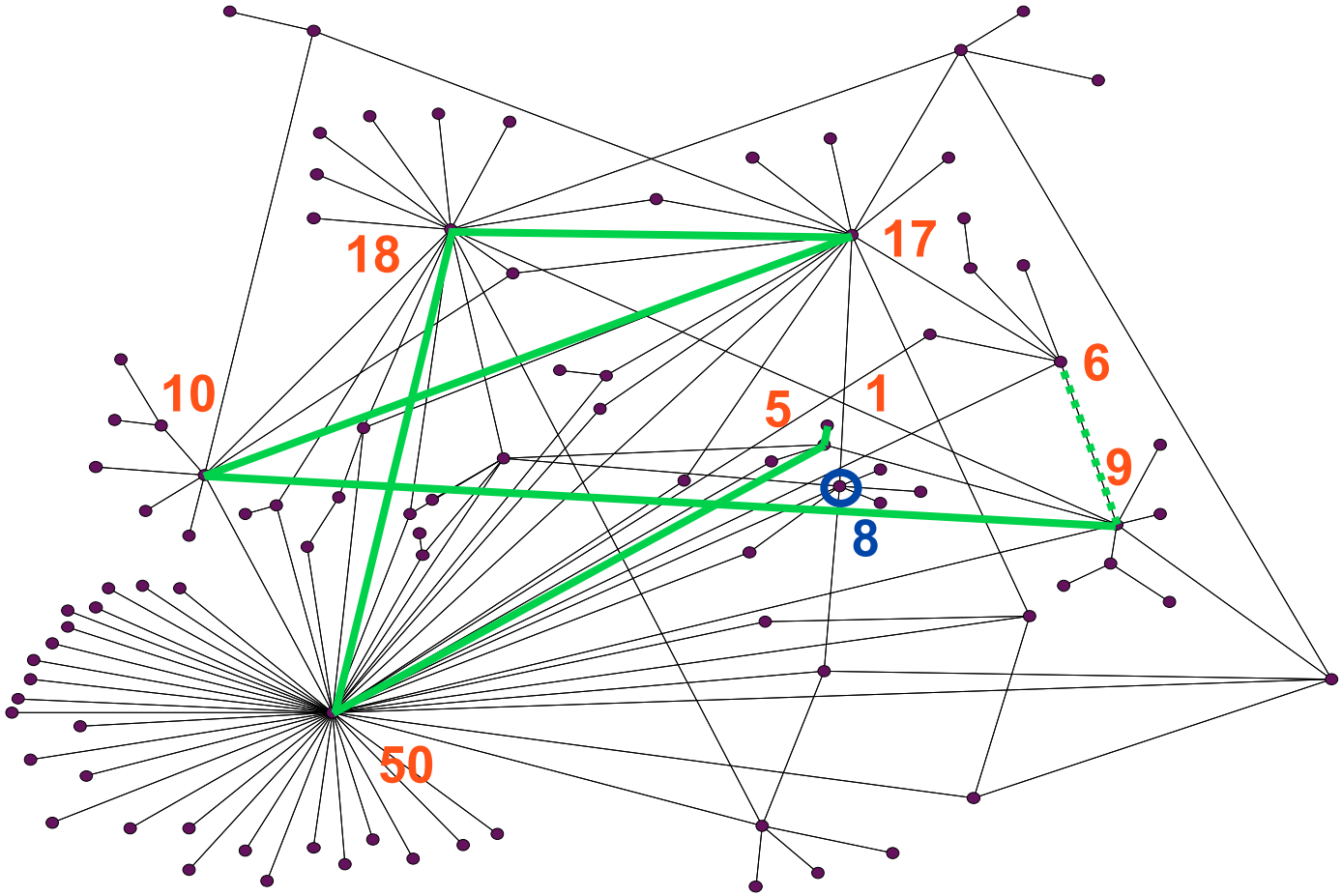
Social networks, AT&T call graph $\tau \sim 2.1$

Actor collaboration graph
(imdb database)

$\tau \sim 2.0\text{-}2.2$



Following the degree sequence

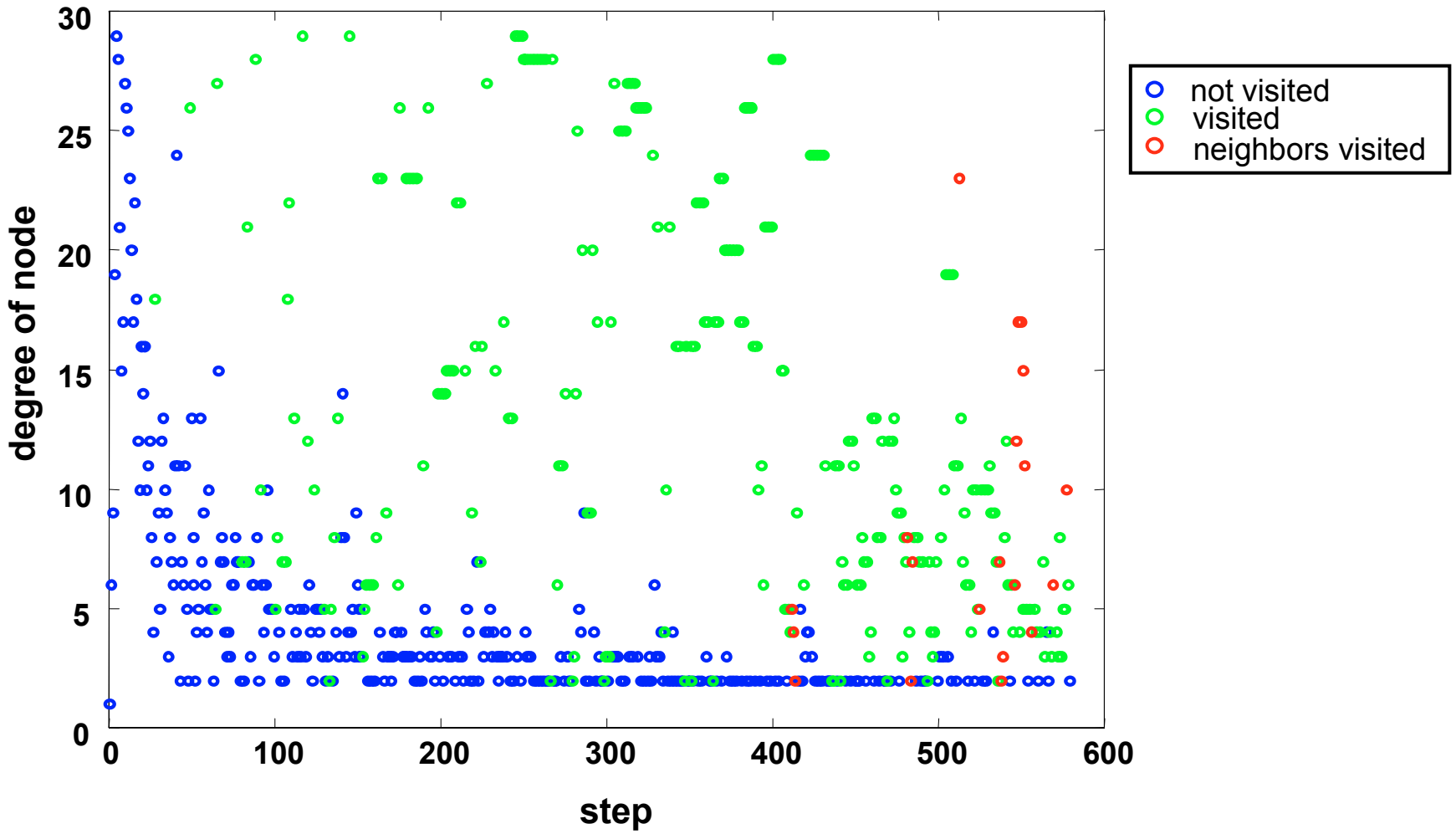


Complications

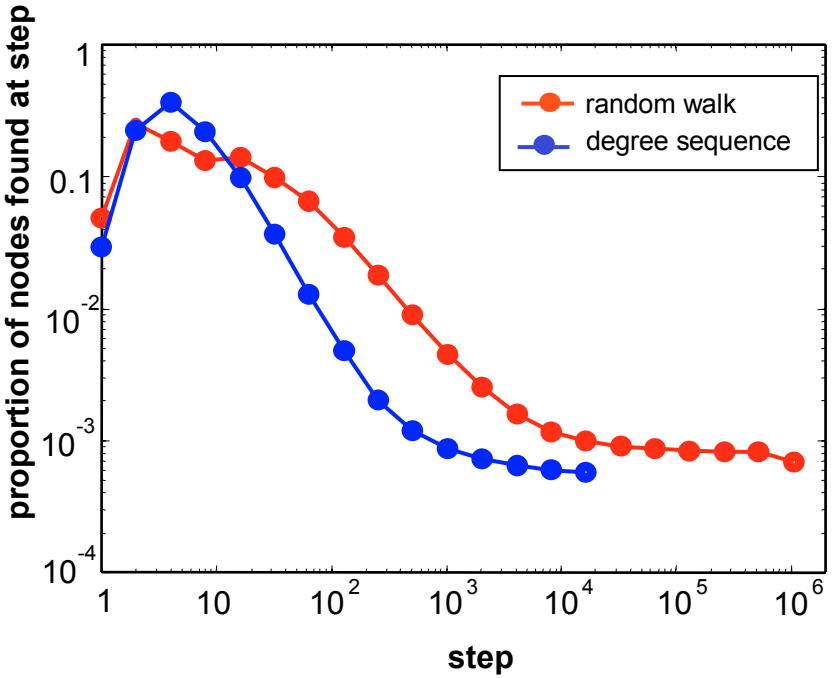
Should not visit same node more than once

Many neighbors of current node being visited were also neighbors of previously visited nodes, and there is a bias toward high degree nodes being ‘seen’ over and over again

Status and degree of node visited

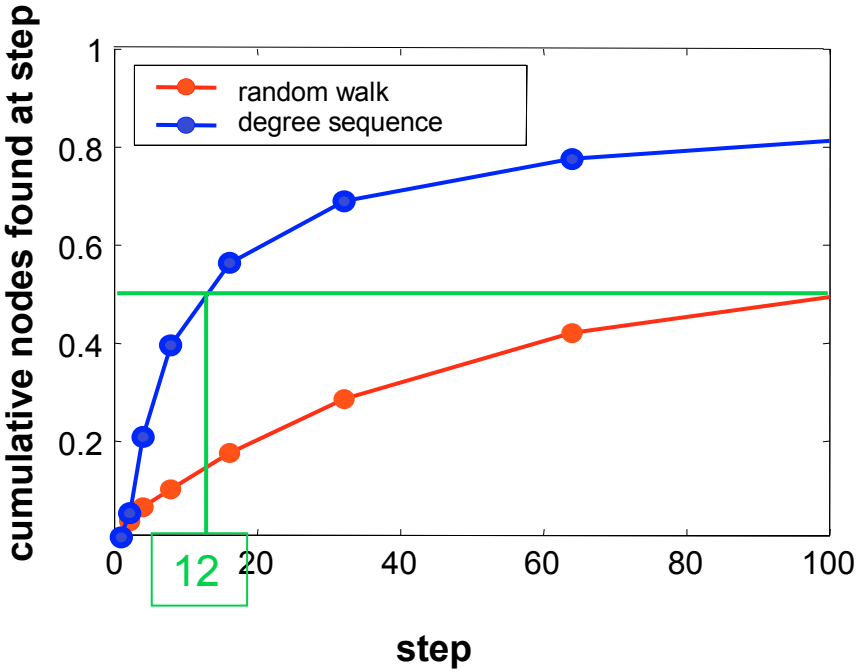


Progress of exploration in a 10,000 node graph knowing 2nd degree neighbors

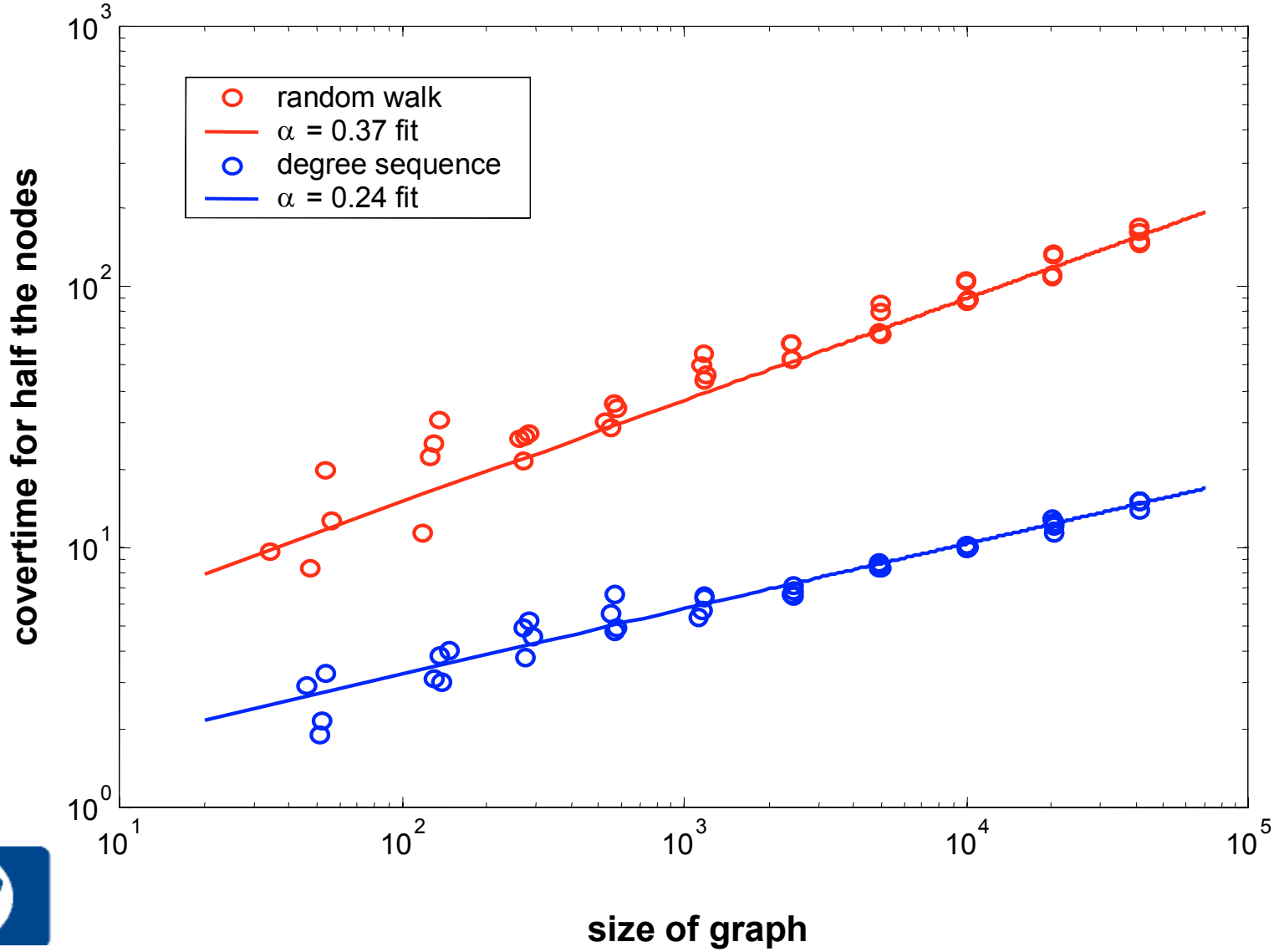


about 50% of a 10,000 node graph is explored in the first 12 steps

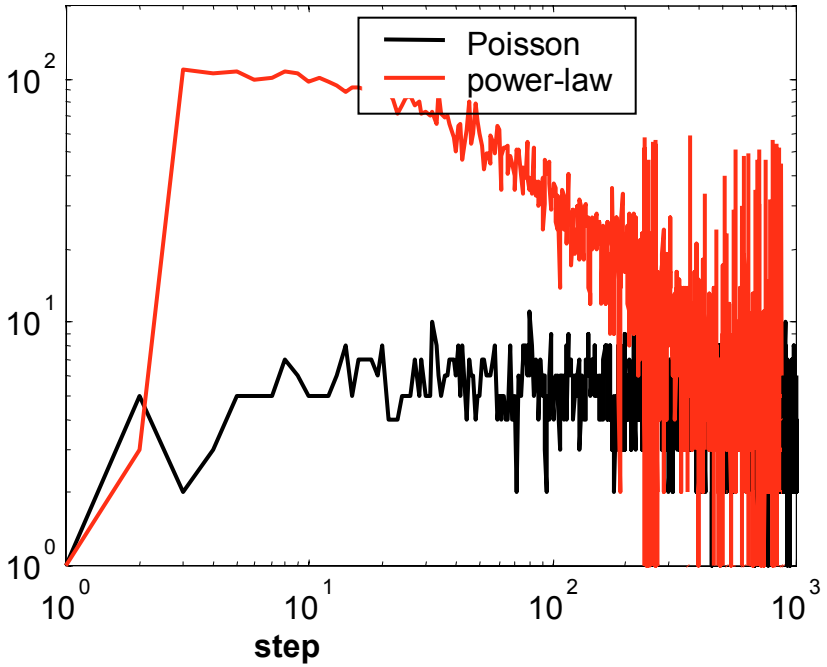
seeking high degree nodes speeds up the search process



Scaling of search time with size of graph



Comparison with a Poisson graph

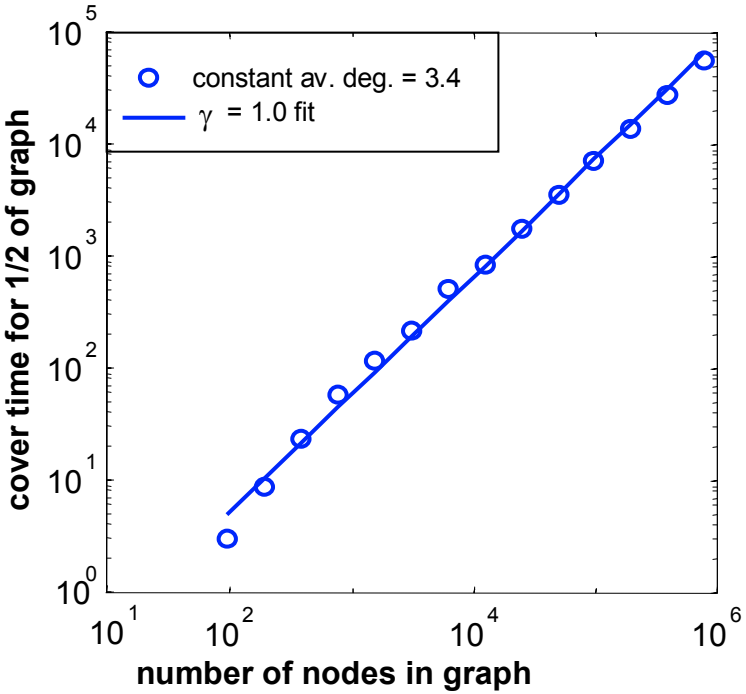


$$G_0(x) = e^{z(x-1)}$$

$$G_1(x) = \frac{x}{z} G_0'(x) \; = G_0(x)$$

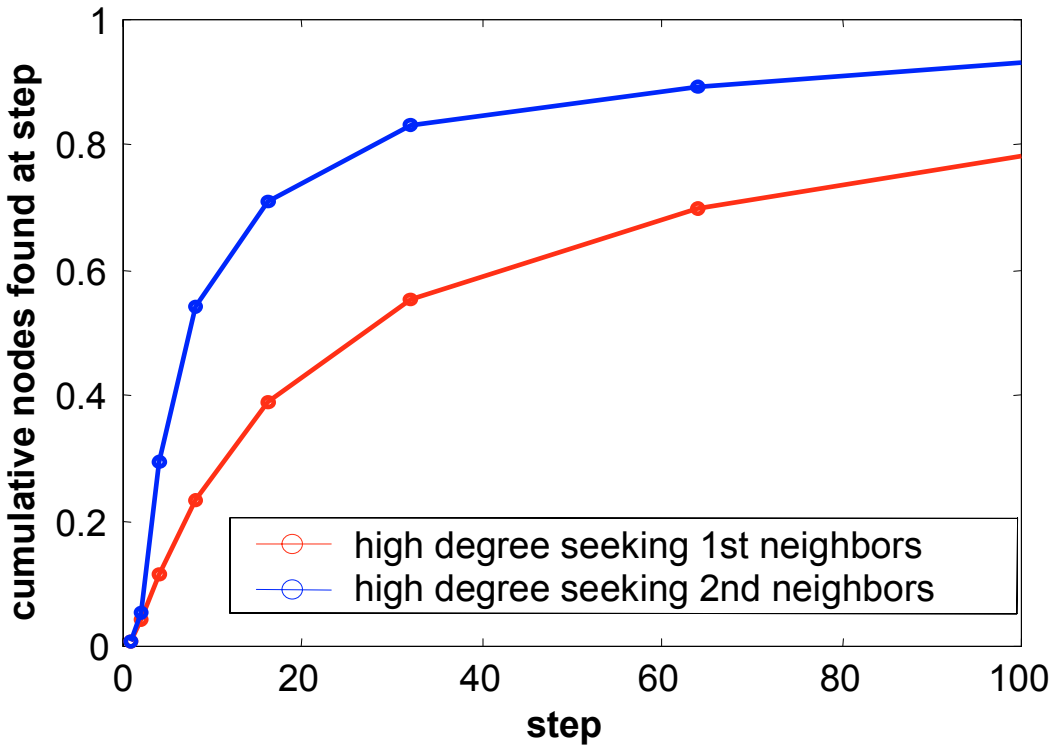
expected degree and expected degree following a link are equal

scaling is linear



Gnutella network

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



Required modifications to nodes

- Maintain a list of files in their neighborhood
- Check query against list.
- Periodically contact neighbors to maintain list
- Append ID to each query processed

Tradeoff

storage/cpu (available) *for* bandwidth (limited)

Theory vs. reality:

- overloading high degree nodes

  but no worse than original scenario where all nodes handle all traffic

  assume high degree -> high bandwidth
  so can carry the traffic load

- fewer nodes used for routing,
  ⇒ system is more susceptible to malicious attack

Partial implementation:

- localized indexing
- traffic routed to high degree nodes

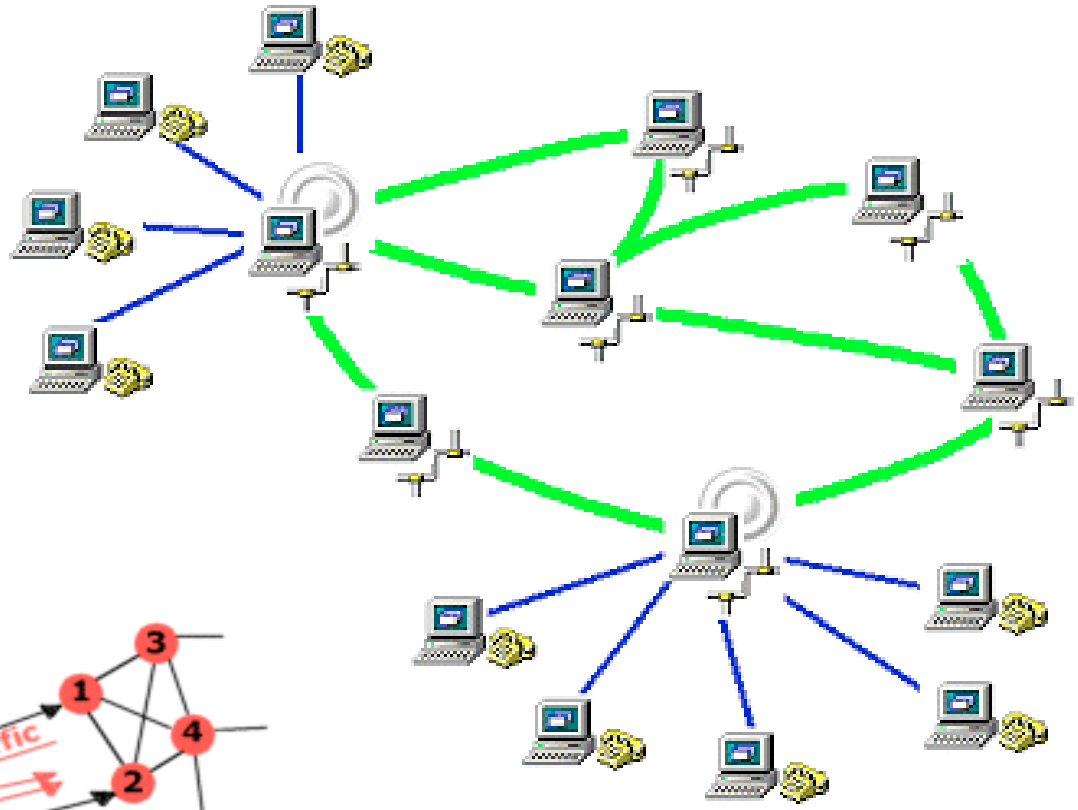
Clip2 Distributed Search Solutions

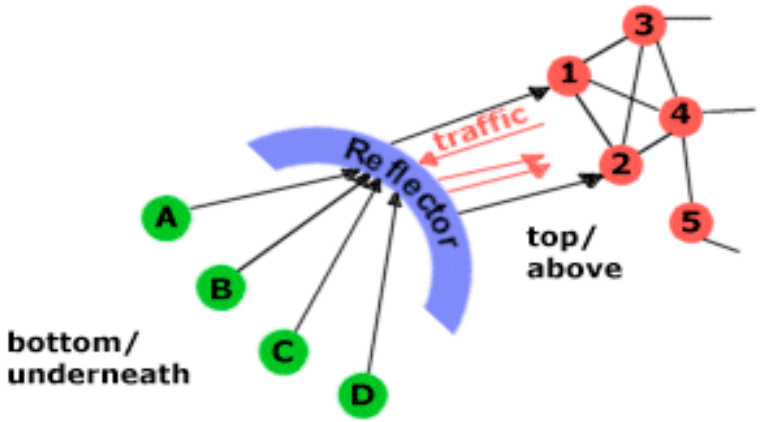
http://dss.clip2.com

Broadband user running Reflector

Broadband user running Gnutella

Dial-up user running Gnutella





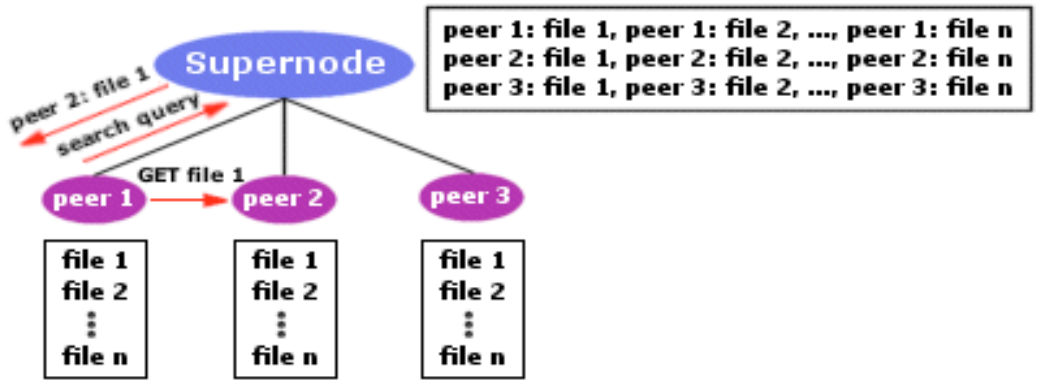
Connection-preferencing rules

LimeWire, BearShare:

drop connections to unresponsive hosts
drives slower hosts to have fewer connections &
move to edge of network

Supernodes

Kazaa, BearShare defender, Morpheus SuperNodes



from Clip2: Morpheus out of the Underworld
http://www.openp2p.com/pub/a/p2p/2001/07/02/morpheus.html

Conclusions

Search is faster and scales in power-law networks

Networks intended to be searched, such as Gnutella,
have a favorable P-L topology

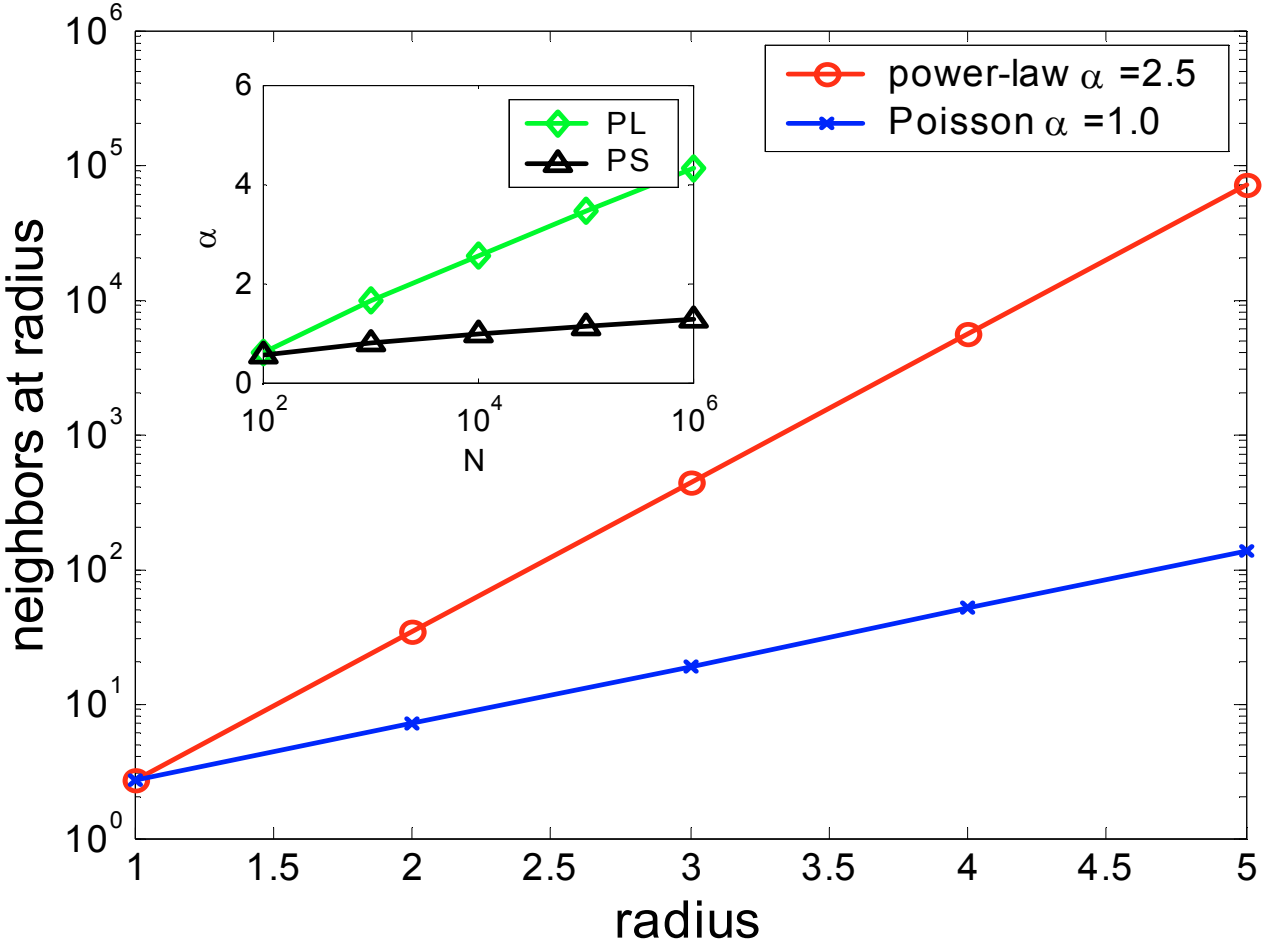
High degree strategy has partially been implemented in existing p2p clients, such as BearShare, Kazaa & Morpheus

A PL link distribution shortens the average shortest path

$$z_r = \alpha^{r-1} z_1 = \left[\frac{z_2}{z_1}\right]^{r-1} z_1$$

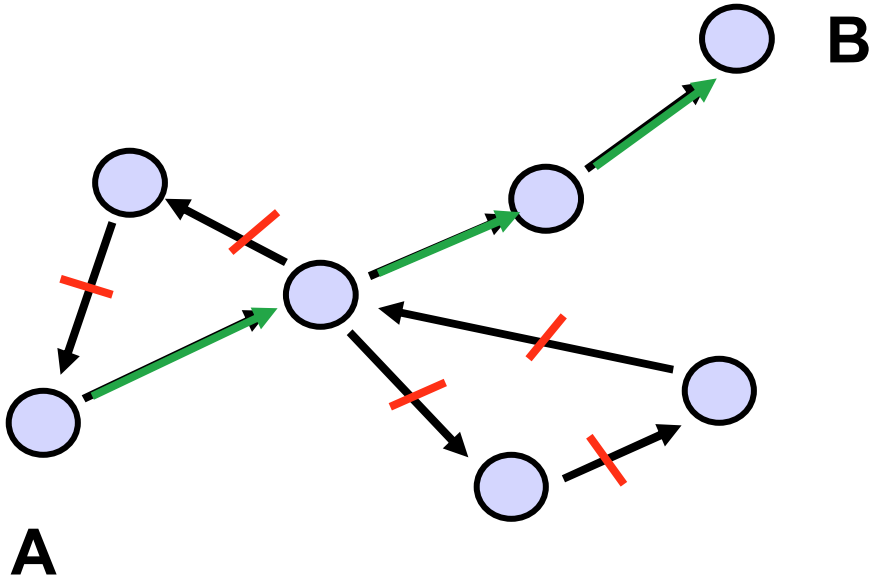
Poisson: $\alpha = z_1$

PL: $\alpha > z_1$


power-law α =2.5
Poisson α =1.0
neighbors at radius
radius
PL
PS
α
N

What about the shortest path discovered along the way?

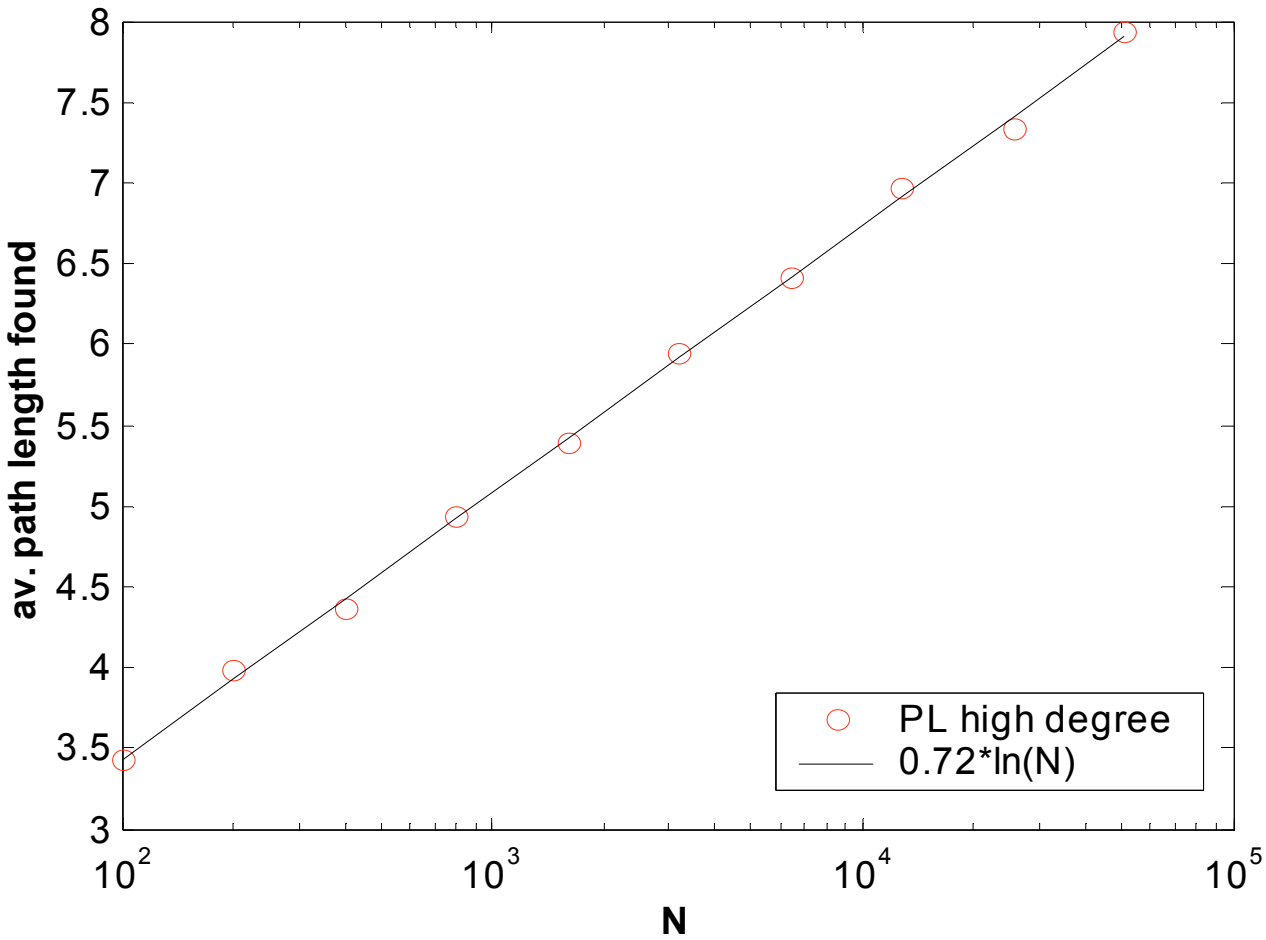
B.J. Kim et al. ‘Path finding strategies in scale-free networks’, PRE (65) 027103.



each node passes message to highest degree neighbor it hasn’t
passed the message to previously

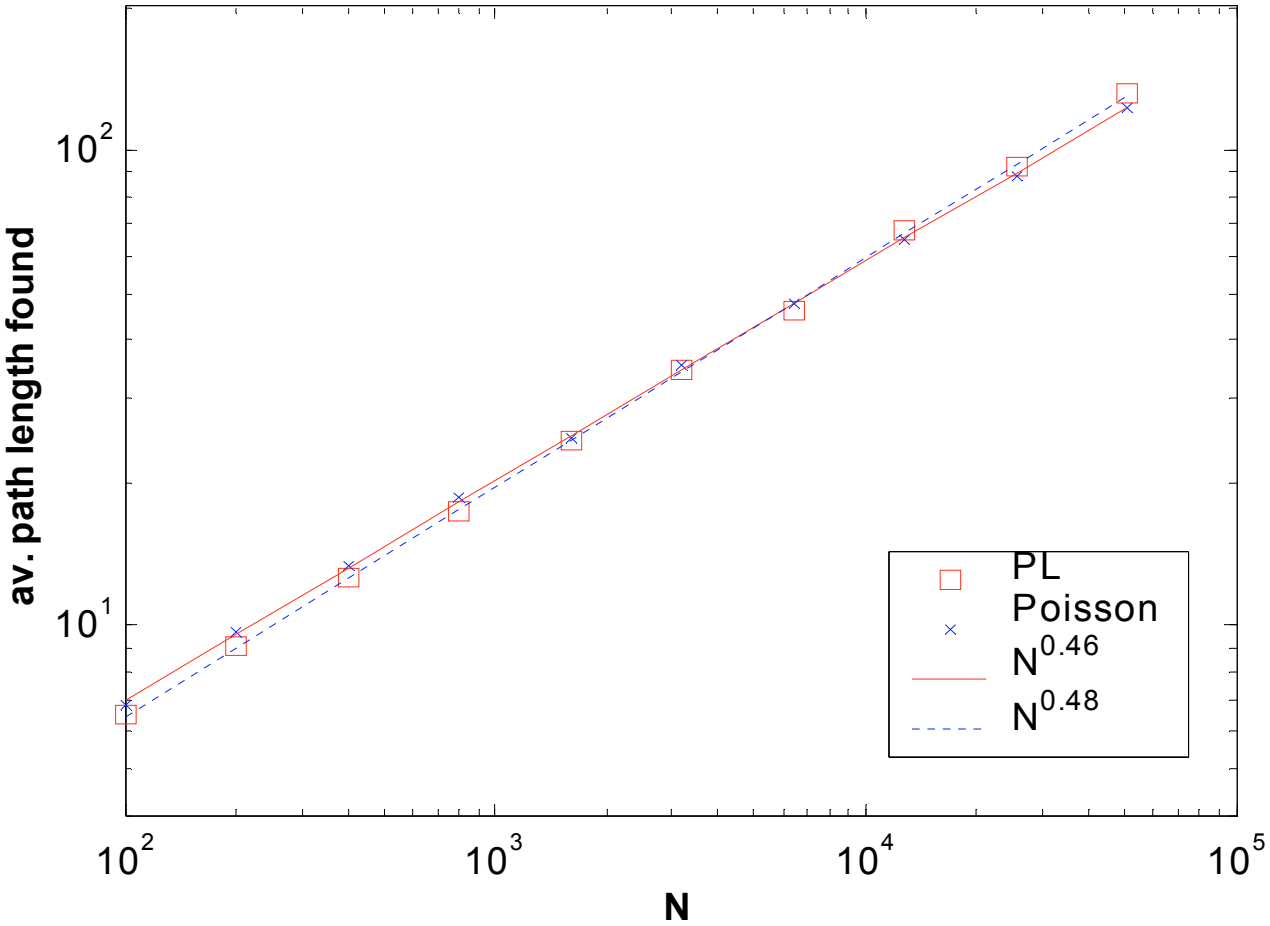
‘cut off’ loops

A high degree seeking strategy finds shortest paths whose average scales logarithmically with the size of the graph

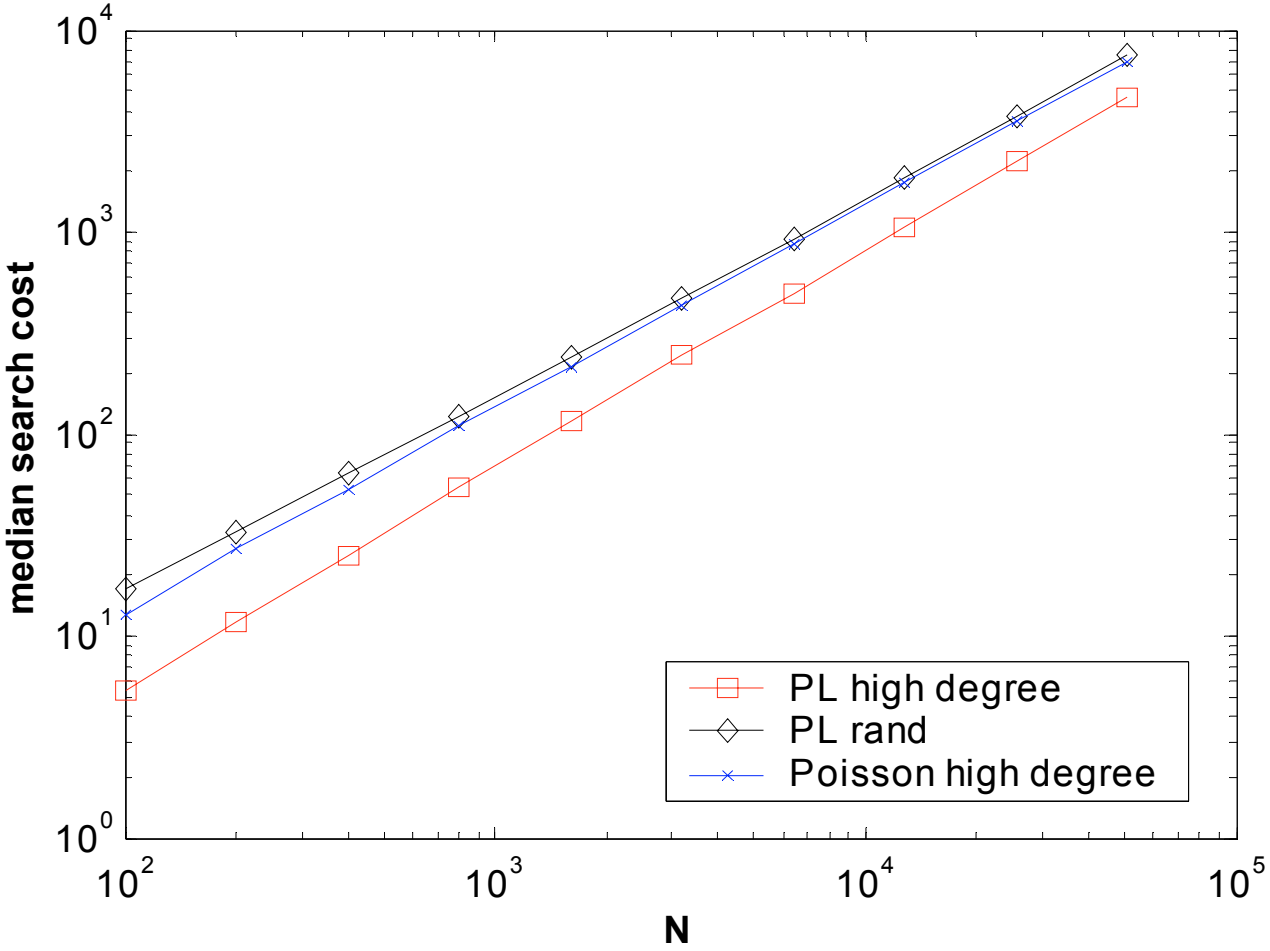


Scaling of the path length found using a

- random strategy on a PL graph
- high-degree strategy on a Poisson graph



But…
Search costs are prohibitive, might as well do a BFS



Freenet

Queries are passed to one peer at a time.

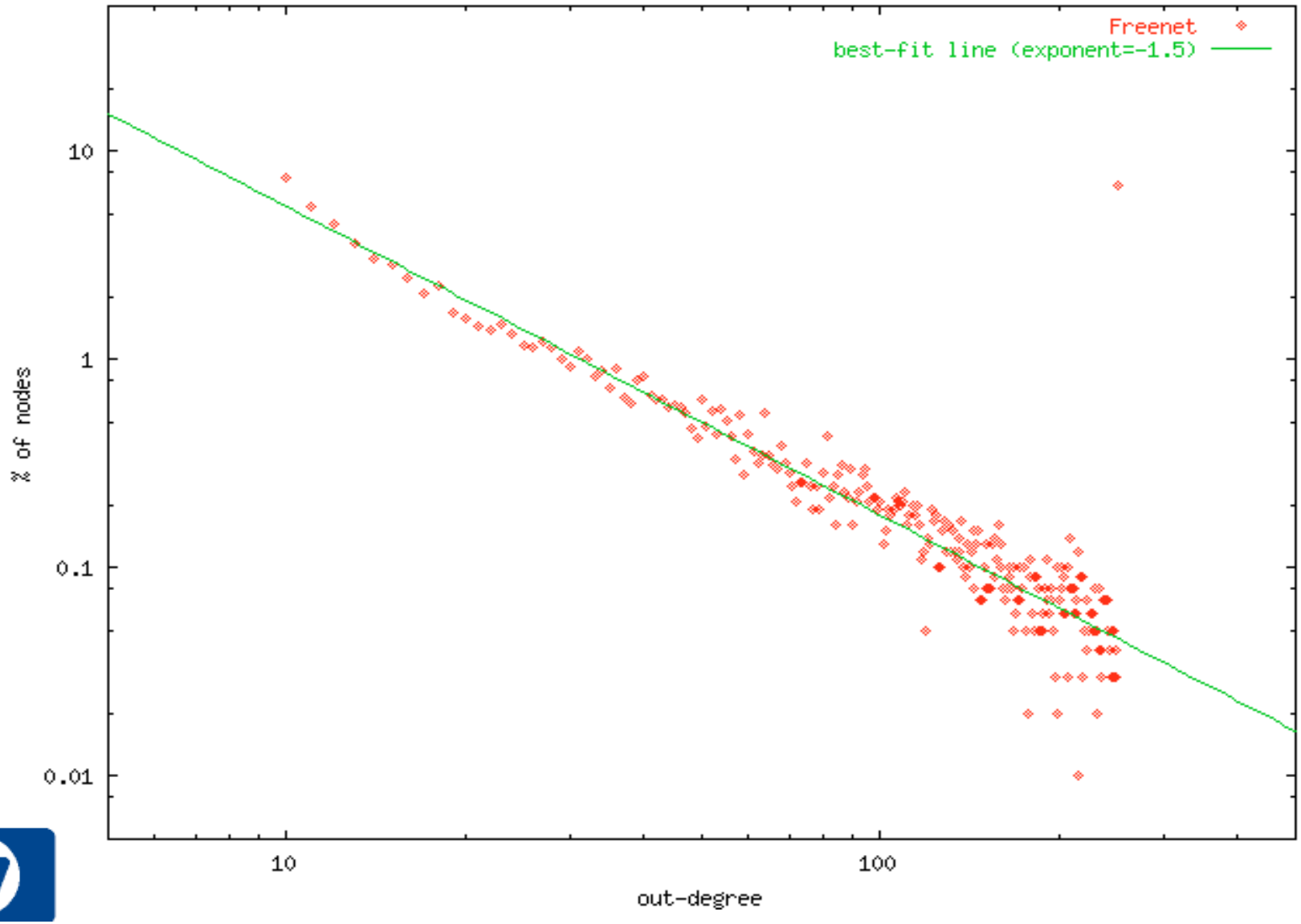
Queries routed to high degree nodes.

Has a power-law topology

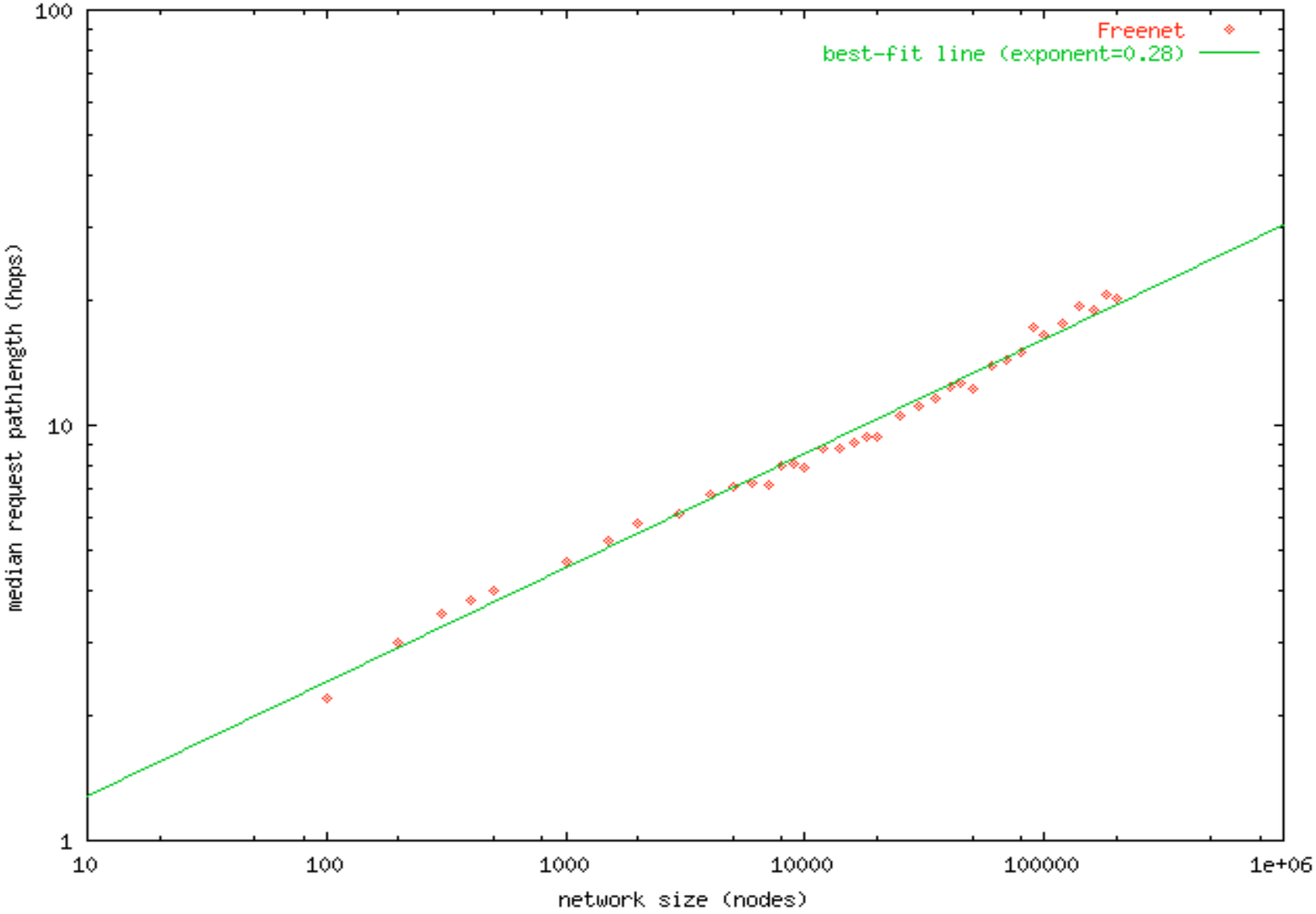
Theodore Hong, 'Performance' chapter in O'Reilley's "Peer-to-Peer, Harnessing the Power of Disruptive Technologies"

Scales as $N^{0.275}$ with the size of the network, N.

Theodore Hong,
power - law link distribution of a simulated Freenet network



Theodore Hong, scaling of mean search time on a simulated Freenet network

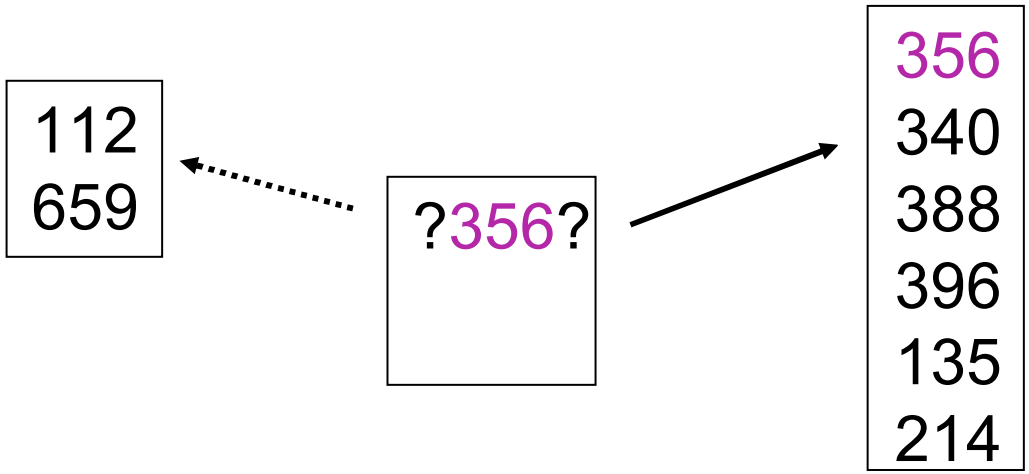


Node specialization key to Freenet's speed

Each node forwards query to node with "closest" hash key

Node passing back a match remembers the address the data came from

Results in nodes developing a bias towards a part of the keyspace



Queries are naturally routed to high degree nodes

Use keys for orientation

To find out more

Information dynamics group at HP Labs
http://www.hpl.hp.com/research/idl

Adamic, Lukose and Huberman,
“Local Search in Unstructured Networks”,
http://www.hpl.hp.com/research/idl/papers/review/