HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Overlay and P2P Networks

On power law networks

Prof. Sasu Tarkoma

3.10.2011

On Zipf's distribution and power-laws

A power-law implies that small occurrences are extremely common, whereas large instances are extremely rare.

This regularity or law is also referred to as Zipf or Pareto

Zipf is used to model the **rank** distributions, and power-law for **frequency** distributions

Examples

- Word popularity rank in English
- Node degree distribution in a network

Zipf's Law

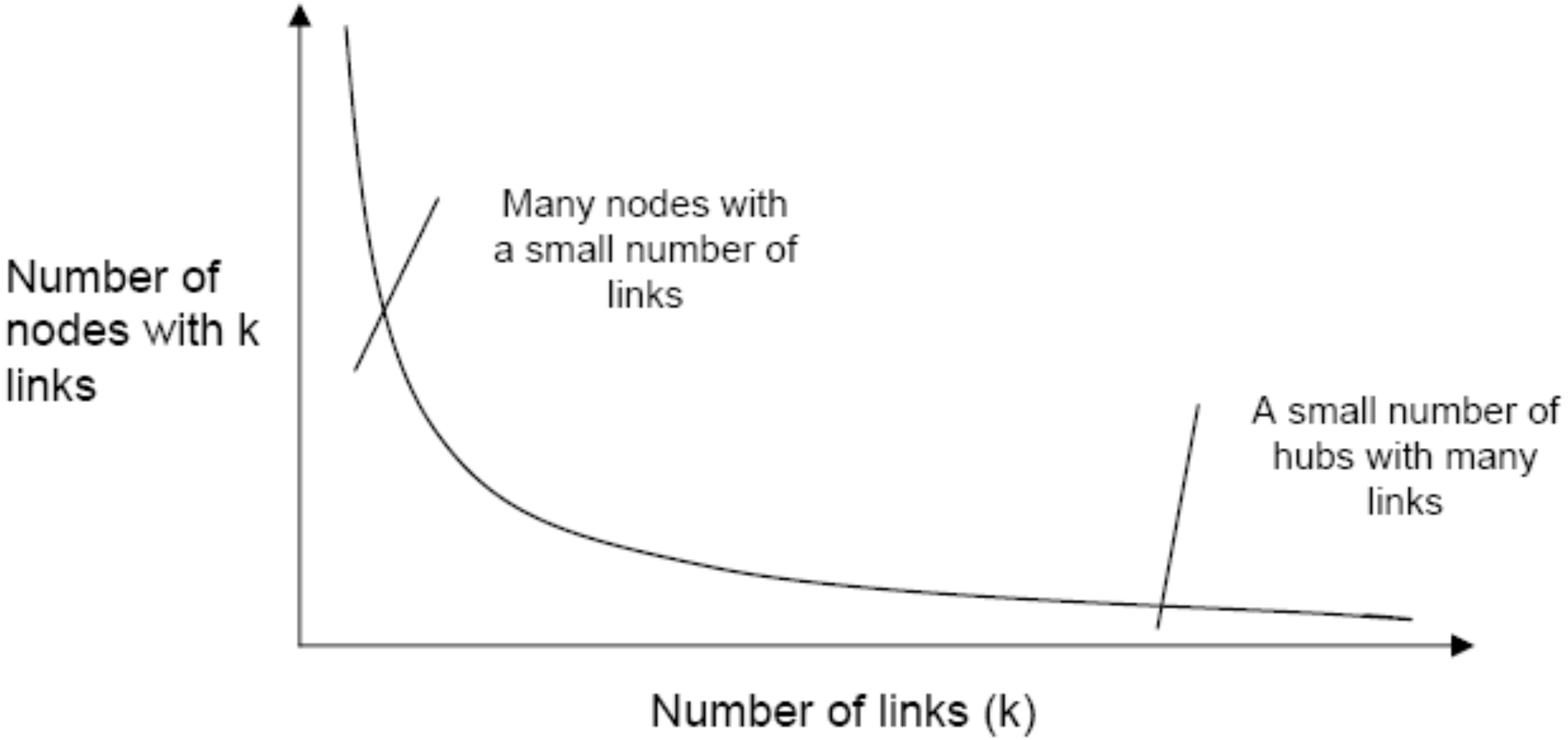
$F \sim R^{-\beta}$, where R is the rank and the constant is close to one

Straight line on a log-log plot

Zipf's law has been used to model Web links and media file references. It has therefore profound implications for content delivery on the Internet.

Efficient caching relies heavily on Zipf's law to replicate a small number of immensely popular files near the users.

Power-law distribution



Applications

The linguist George Zipf first proposed the law in 1935 in the context of word frequencies in languages.

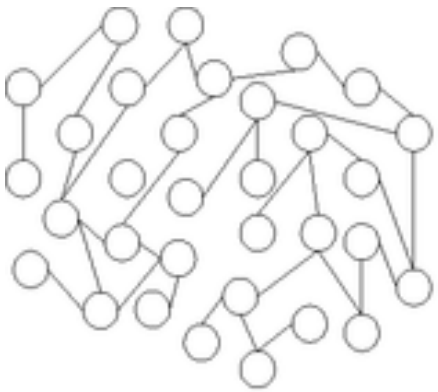
Many applications, for example size of cities, income distributions

No typical scale hence **scale-free** (consider height of people which is Gaussian, no order of magnitude differences)

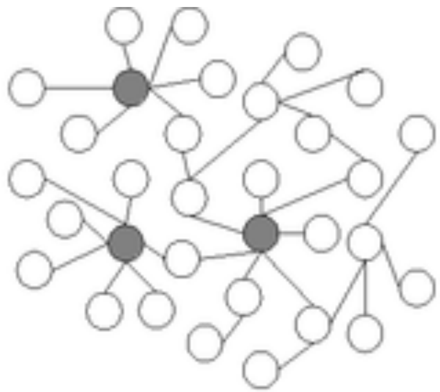
For Web sites, the Zipf law means that large sites get disproportionately more traffic than smaller sites.

Popularity of files follows the distribution → implications for caching

Scale-Free Networks



(a) Random network



(b) Scale-free network

Internet Connections

The distribution of the number of connections a host has to other hosts on the Internet has been shown to follow the power law distribution.

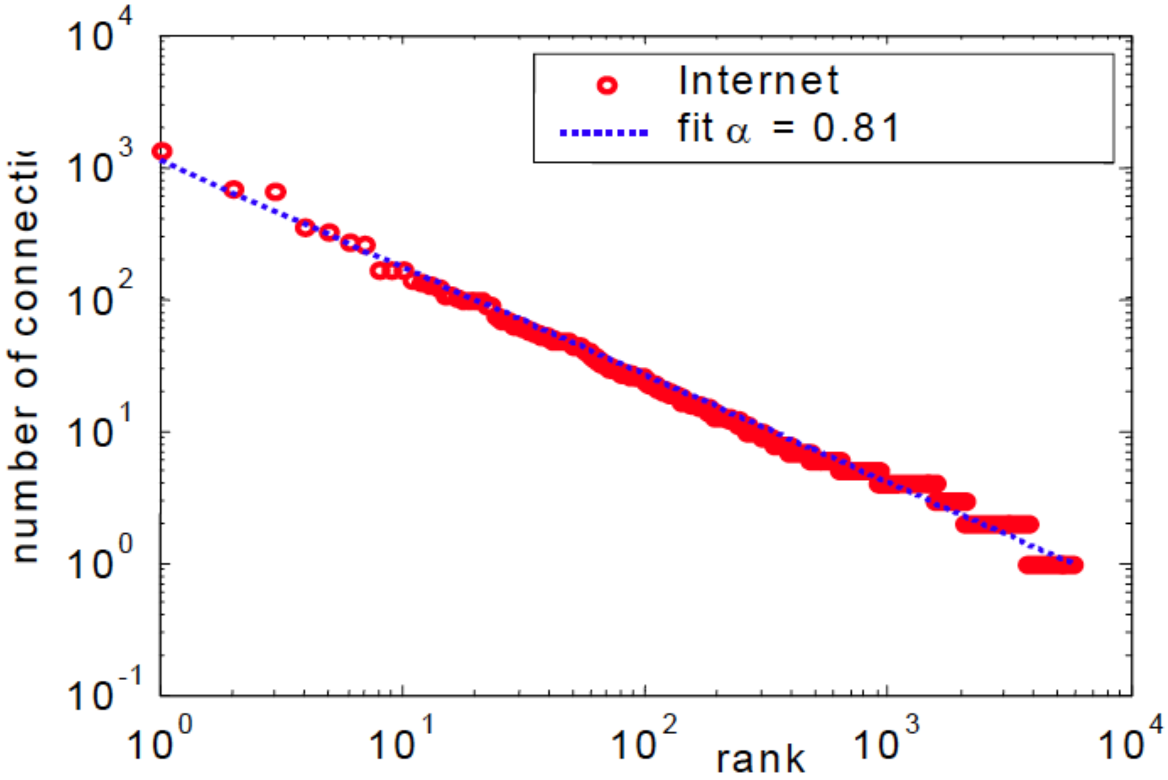
Implications for the P2P algorithms.

Some nodes maintain majority of the connections (the hubs)

Therefore send queries toward hubs.

High-degree nodes may make the network vulnerable to attacks

Example: AS Connectivity



Source: http://www.hpl.hp.com/research/idl/papers/ranking/
adamicglottometrics.pdf

Observations

Gnutella (v0.7) and Freenet support the formation of hubs

They are power law networks

How robust are these networks?

Robustness

Given a certain expected network structure, a very interesting question is how easy it is to disrupt the network and partition it into disjoint parts.

Cohen et al. have analytically shown that networks in which the vertex connectivity follows a power-law distribution with an index of at most (alpha<3) are very robust in the face of random node breakdowns.

$$p \leq 1 + \left(1 - m^{\alpha-2} K^{3-\alpha} \frac{\alpha - 2}{3 - \alpha}\right)^{-1},$$

where p is a probability bound on network partitioning, m is the minimum node degree, and K is the maximum node degree.

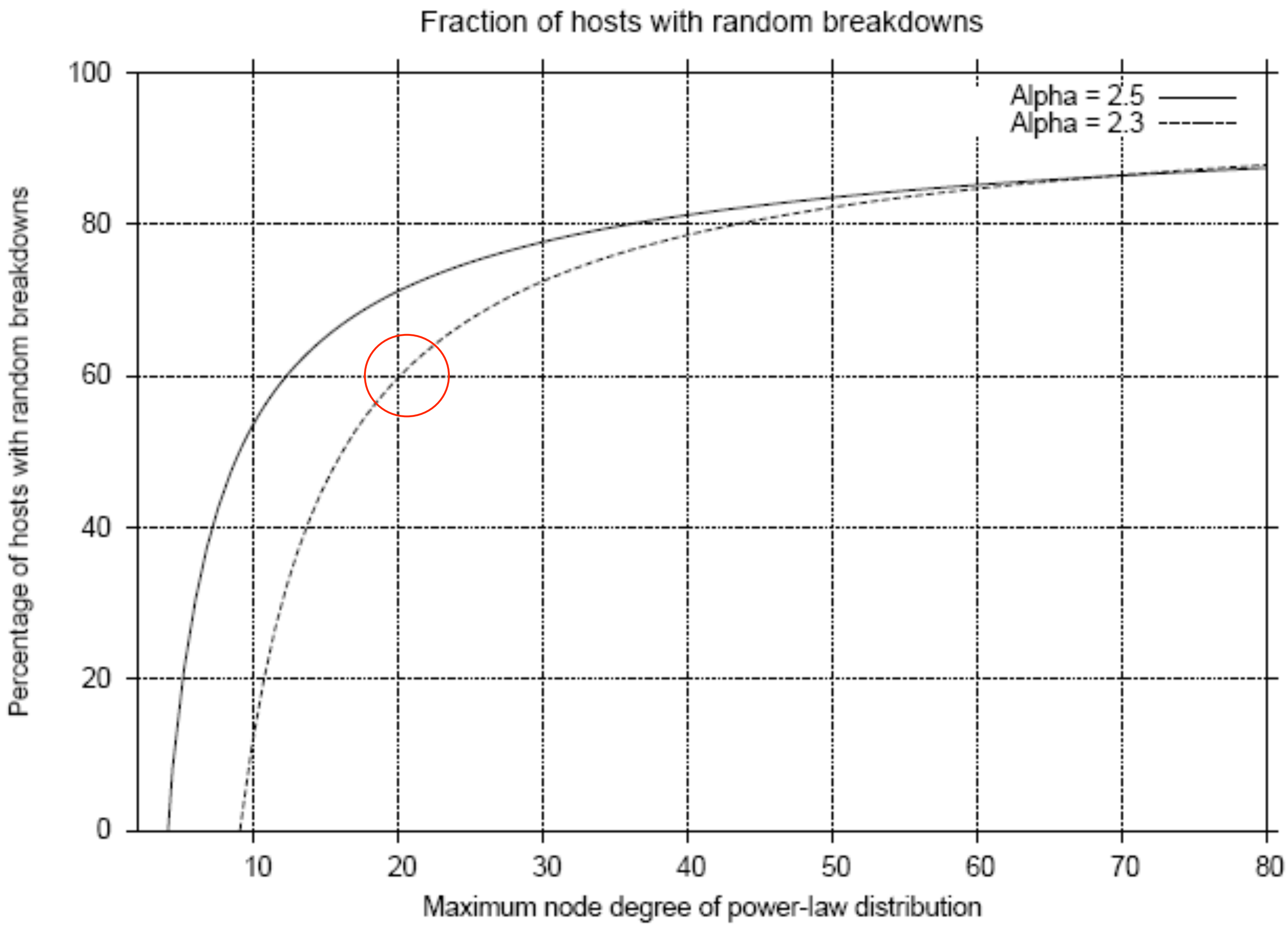
Robustness II

The Internet node connectivity has been shown to follow a power-law distribution with alpha=2.5.

Similar investigation has been made for the Gnutella P2P network resulting in the observation that alpha = 2.3

Both the Internet and Gnutella present a highly robust topology. They are able to tolerate random node breakdowns.

Resiliency of power-law networks



Gnutella Robustness

For a maximum and fairly typical node degree of 20, the Gnutella overlay is partitioned into disjoint parts only when more than 60% of the nodes are down.

Robustness is a highly desirable property in a network. The above equation is useful in understanding the robustness of power-law networks; however, it assumes that the node failures are random.

Orchestrated Attacks

Although a power-law network tolerates random node failures well, it is still vulnerable to selective attacks against nodes.

An orchestrated attack against hubs in the network may be very effective in partitioning the network.

Small Worlds: Milgram's experiment

The Small-World Problem – Milgram (1967).

How many intermediaries are needed to move a letter from person A to person B through a chain of acquaintances?

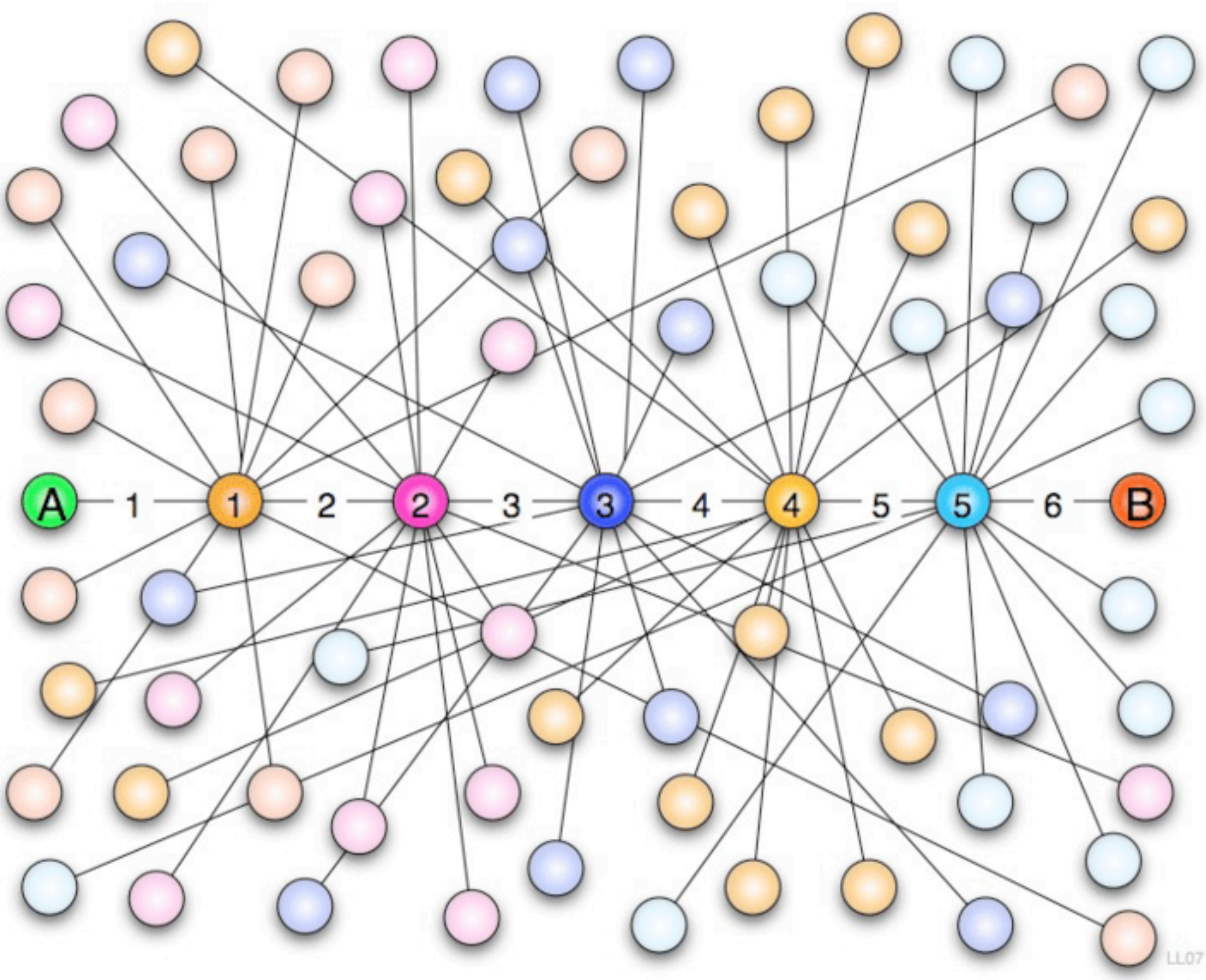
Designed to find out average path length.

Letter-sending experiment: starting in Nebraska/Kansas,with a target person in Boston.

People forwarded the message towards the target person.

Six degrees of separation.

Six degrees of separation



Source: Wikipedia

Small Worlds

Small-world networks are characterized by a **graph degree power-law distribution (Finnish: potenssijakauma).**

Definition: A small world network is a network with a dense local structure and a diameter comparable to a random graph

Also the term **scale-free** is used for these networks.

They exhibit **clustering** and thus are different from random networks (preferential attachment).

Most nodes have relatively few local connections to other nodes, but a significant small number of nodes have large wide-ranging sets of connections.

Barabási-Albert Model

Scale-free networks with power-law node degree distribution

The network grows in time

No random edge generation, higher the degree, higher the probability that the new vertex will attach (preferential attachment)

Local Clustering Coefficient

The clustering coefficient C(v) of vertex v in a directed graph is given by

the number of **links between the vertices within its neighborhood** divided by the number of **links that could possibly exist between them**

Neighbourhood is the immediately connected neighbours

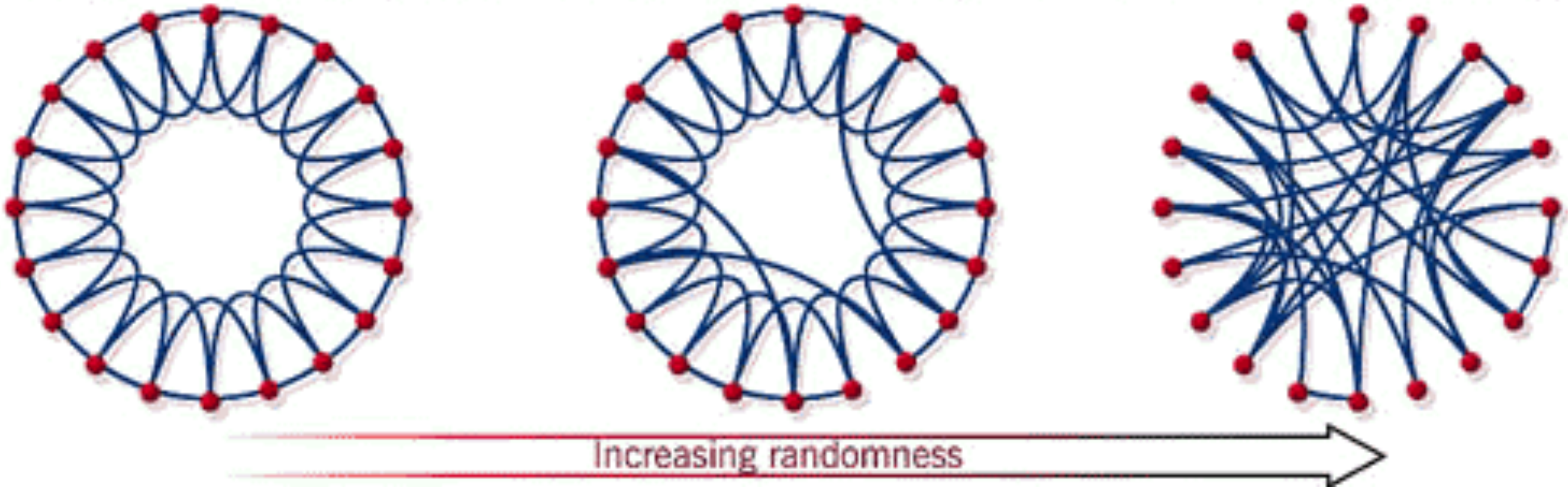
k(k-1) possible links for k vertices

Network average: $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i.$

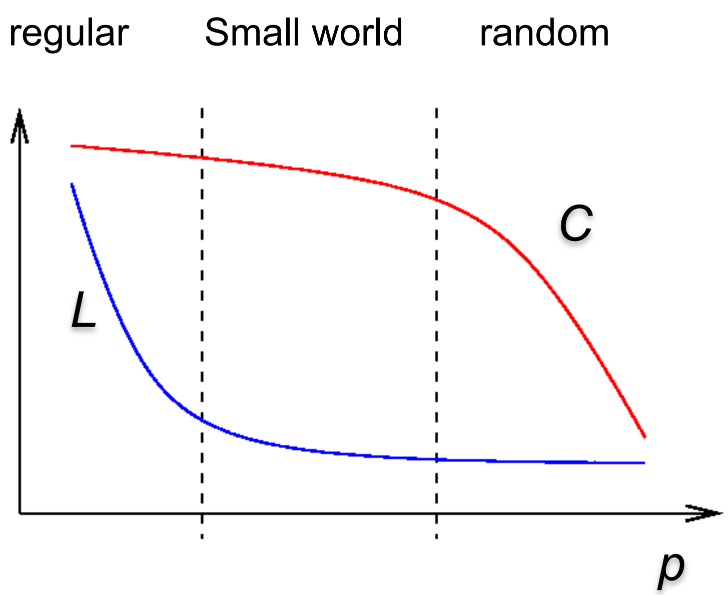
Structured network
- high *clustering*
- large diameter
- regular

Small-world network
- high *clustering*
- small diameter
- almost regular

Random network
- small *clustering*
- small diameter



Reference: Duncan J. Watts & Steven H. Strogatz, Nature 393, 440-442 (1998)



C(p) : clustering coefficient
L(p) : average path length

Reference: Duncan J. Watts & Steven H. Strogatz, Nature 393, 440-442 (1998)

Kleinberg’s result

Jon Kleinberg showed that it is possible to do efficient routing on grids with the small world property.

The possibility of efficient routing depends on a balance between the proportion of shortcut edges of different lengths with respect to coordinates in the base grid.

The key idea is to use a frequency of edges of different lengths that decrease **inverse proportionally** to the length.

Kleinberg’s result II

Results in an infinite family of small world network models on a grid with power-law distributed random long-range links

K(n,k,p,q,r)
p – radius of neighbours to which short local links
q – number of random long range links
k - dimension of the mesh
r - clustering exponent of inverse power-law distribution.

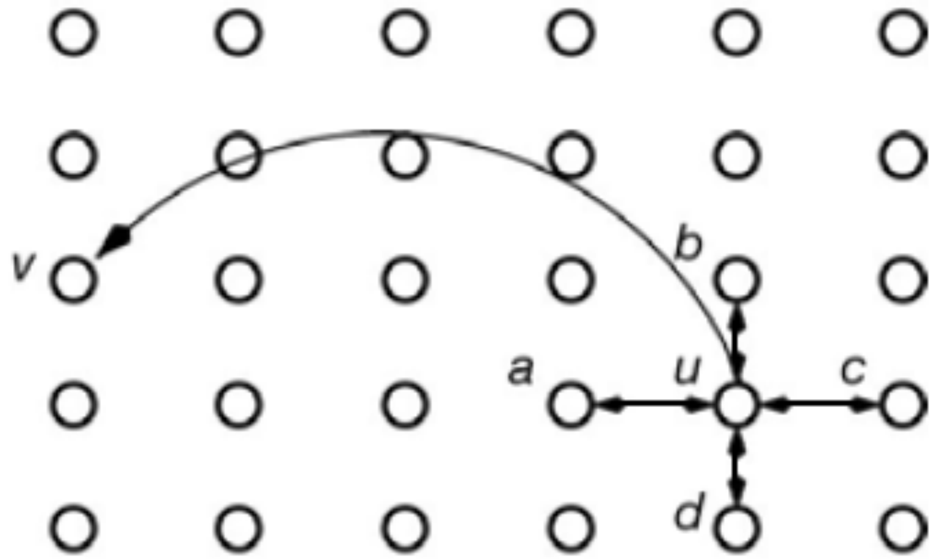
Prob.[(x,y)] $\propto$ dist(x,y)$^{-r}$

Expected Delivery time =

$O((\log n)^2)$, for r = 2 (and the special case k=r).
$\Omega(n^{(2-r)/3})$, for $0 \leq r < 2$.
$\Omega(n^{(r-2)/(r-1)})$, for $2 < r$.

Example

Node u is connected to all its neighbors (a, b, c , and d) and has a long-range link to some randomly chosen node v with a probablility proportional to dist(u, v)$^{-r}$



Just using the neighbours gives O(n) for destination
If the clustering coefficient is zero, then the long range links are too random
If one then there are too few random links
Two would be the optimal value (links are uniformly distributed over all distances)
Results in logarithmic diameter for the network

Theorem

Theorem: The routing algorithm will find short paths if and only if k = r.

The idea behind the proof is that for any r < k there are too few random edges to make the paths short.

For r > k there are too many random edges, and thus too many choices to which the message could be sent.

Kleinberg’s result III

Simple greedy routing can find routes in $O(\log^2(n))$ hops, where n is the size of the graph
- Decentralized
- Decisions based on local information
- $O(\log^2(n))$ links are needed

Later work has investigated other topologies than grids (rings, …) and improving efficiency through topology information, cues, etc.

Implication of result: greedy and local solution for building peer-to-peer overlay networks!

Freenet Routing Revisited

Every file is has a key (derived via a hash function)

A file is stored at some node with a similar key

At each peer each request is forwarded to the node in its routing table having the closest key to the requested one

If the request is successful, the file is sent back via the routing nodes and each node saves the file and adds the sending node's address to its local routing table (i.e., frequently requested files are replicated)

If the routing table is full, the least recently used (LRU) entry is evicted

Clustering and caching for achieving the small world network benefits in routing

Freenet Idea

Assume that the network exhibit small world properties.

Should be possible to recover an embedded Kleinberg small-world graph.

This is accomplished by selecting random pairs of nodes and potentially swapping them based on an objective function.

Function minimizes the product of all the distances between any given node and its neighbors.

Is Freenet a small world?

There must be a scale-free power-law distribution of links within the network.

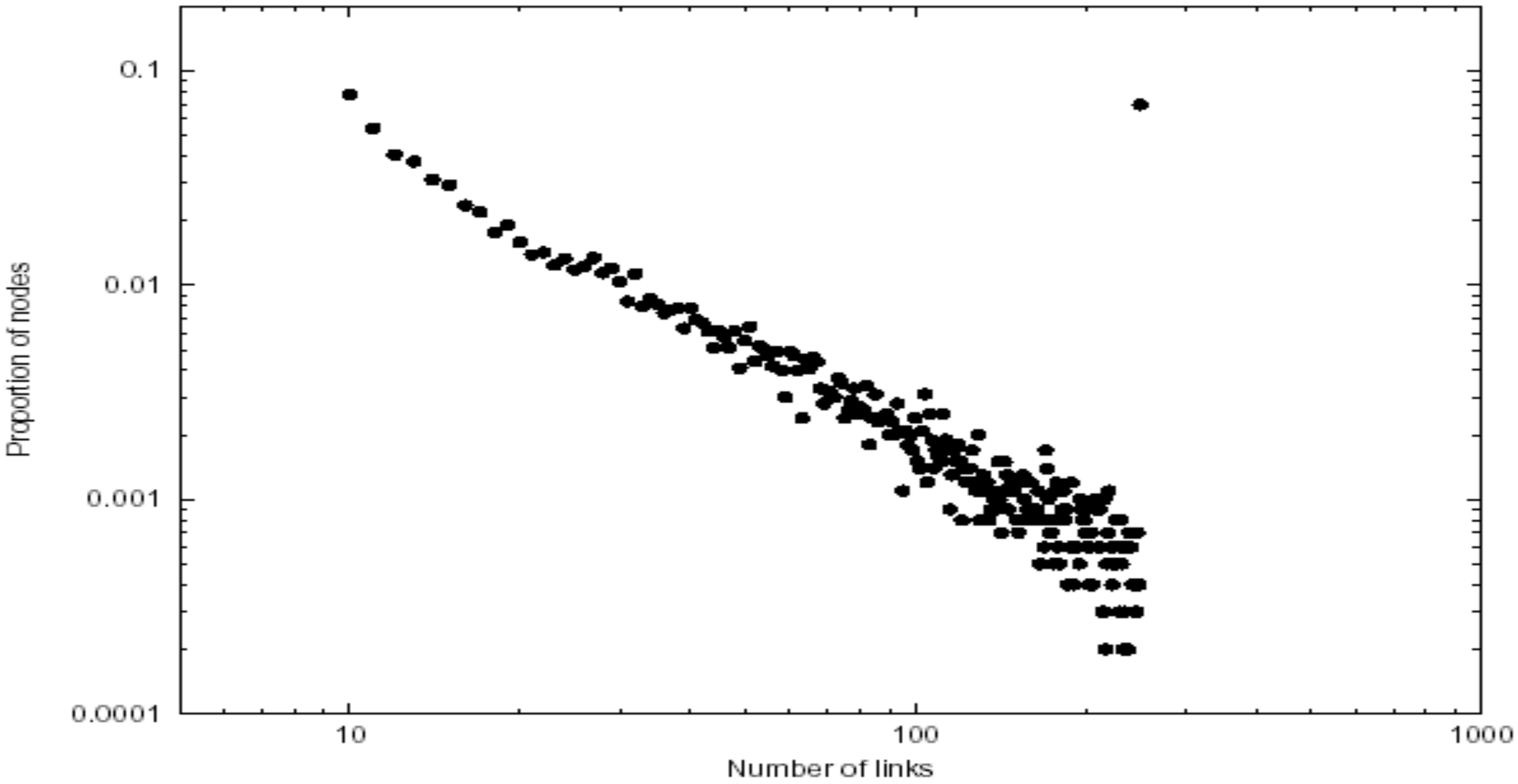


Fig. 5. Distribution of link number among Freenet nodes.

Source: www.ics.forth.gr/dcs/Activities/Projects/p2p/ploumid-freenet.ppt

Applications of Small World Networks

Many applications in peer-to-peer networks

The Gnutella network has been observed to exhibit the clustering and short path lengths of a small world network. Its overlay dynamics lead to a biased connectivity among peers where each peer is more likely connected to peers with higher uptime

The Freenet routing algorithm is built on the small world assumption

Other applications in distributed hashing (DHTs) such as Symphony that uses long-range contacts drawn randomly from a family of harmonic distributions