Peer-to-Peer Networks

Chapter 4: Graphs and Methods

Thorsten Strufe

Note: these slides have been prepared with influence by material of Prof. Michael Welzl, Prof. Pietro Michiardi, and Dr. Stefan Schmid

Classes of Graphs

- Regular graphs
- Random graphs
- Graphs with Small-World characteristic
- Scale-free graphs
- Graphs with plenty more characteristics
  - (Dis-) Assortativity
  - Rich-club connectivity

Regular Graphs

- Regular graphs have traditionally been used to model networks

- Regular graphs:
  - Node degree is constant
  - Different topologies possible

- But the model does not reflect reality of nature very well…

Random Graphs

- Random graphs are first widely studied graph family
  - Many P2P networks choose neighbors more or less randomly

- Two different notations generally used:
  - Erdös and Renyi - ER(n,|E|)
  - Gilbert - G(n,p)

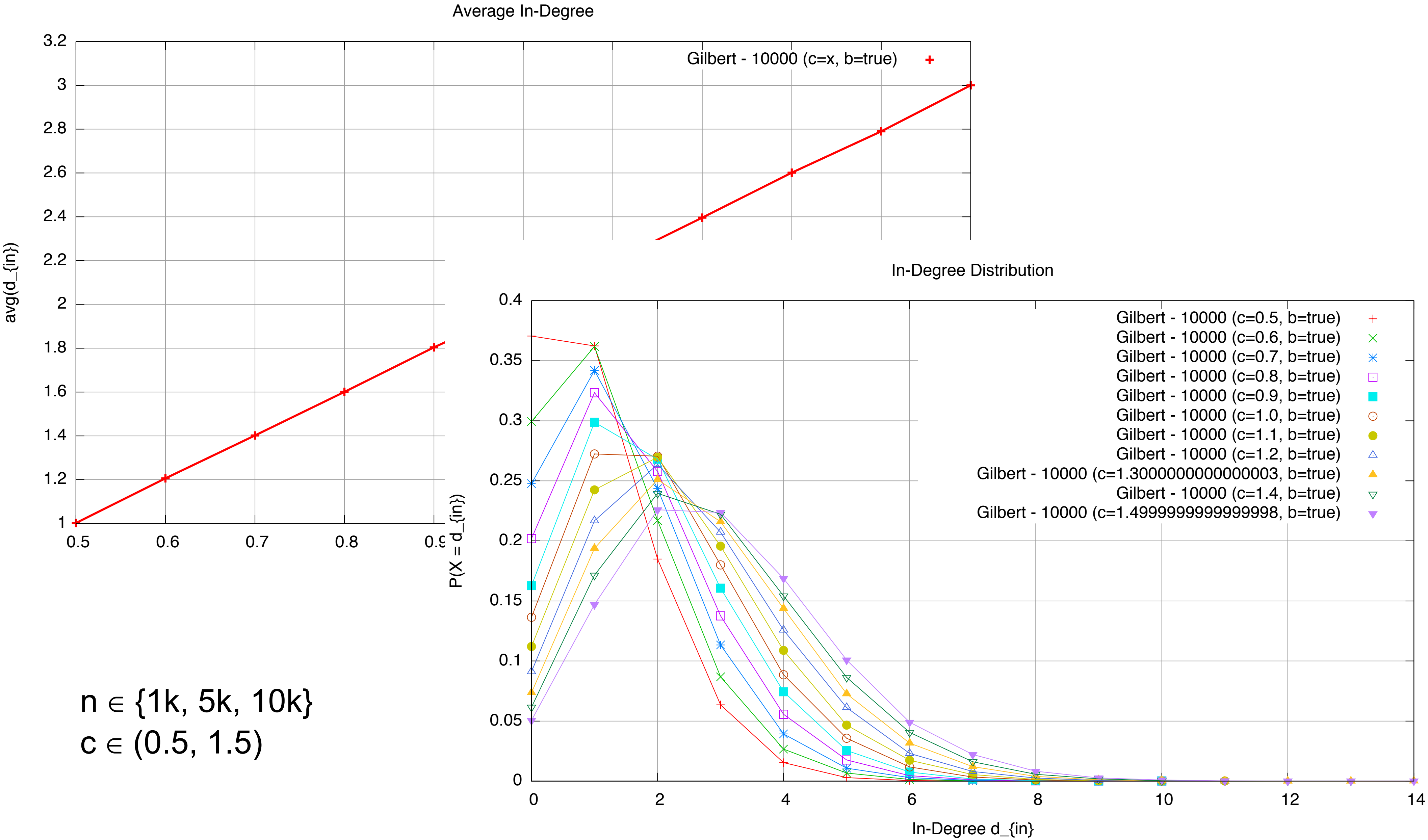
- G(n,p) is a graph where the probability of an edge e = {v, w} is 2p

- Construction algorithm:
  - For each possible edge, draw a random number
  - If the number is smaller than 2p, then the edge exists
  - p can be function of n or a constant

Random Graphs - Basic Properties

- Giant Connected Component
  - Let $c > 0$ be a constant and $p = c/n$.
  - If $c < 1$ every component of $G(n,p)$ has order $O(\log N)$ with high probability.
  - If $c > 1$ then there is one component of size $O(n)$, with high probability.
    - All other components have size $O(\log N)$
  - English: Giant connected component emerges with high probability when average degree is about 1
- Node degree distribution
  - If we take a random node, how high is the prob. $P(k)$ that it has degree k?
  - Node degree is Poisson distributed
  - Parameter c = expected number of occurrences
- Clustering coefficient
  - Clustering coefficient of a random graph is asymptotically equal to p with high probability

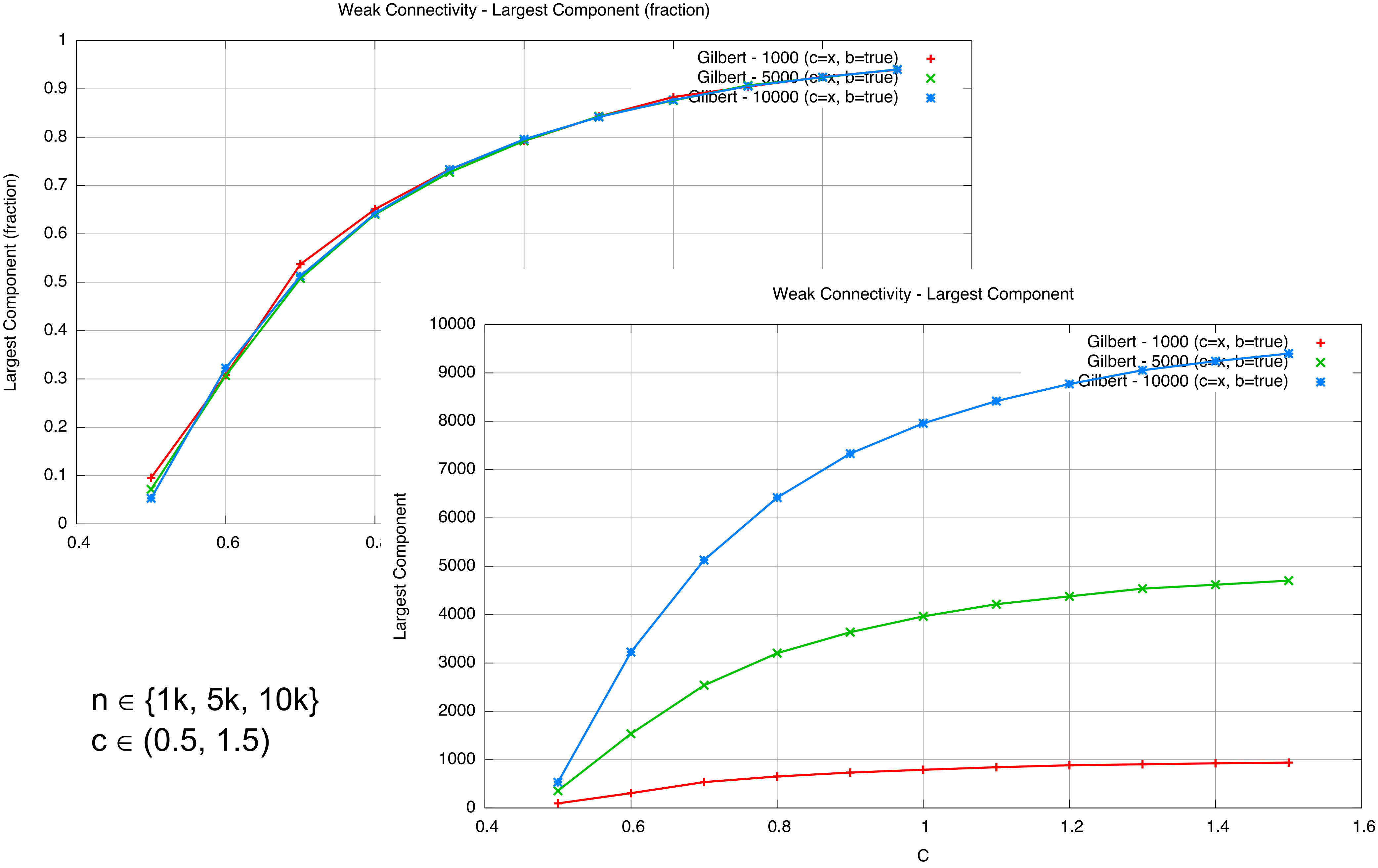
Gilbert - Degree Distribution



$n \in \{1k, 5k, 10k\}$

$c \in (0.5, 1.5)$

Gilbert - Connectivity



n ∈ {1k, 5k, 10k}

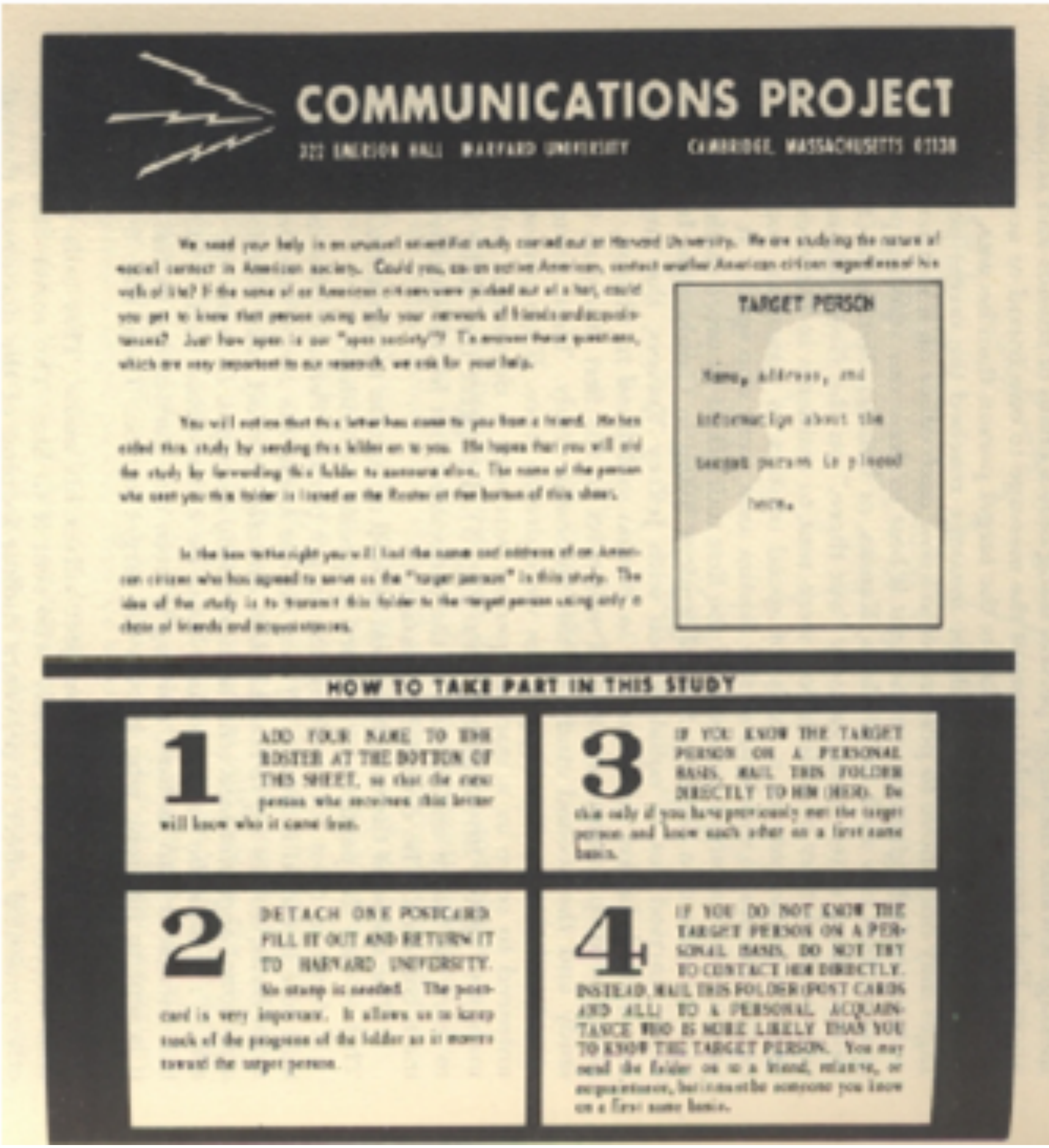
c ∈ (0.5, 1.5)

Random Graphs: Summary

- Before random graphs, regular graphs were popular
  - Regular: Every node has same degree
- Random graphs have two advantages over regular graphs
  1. Many interesting properties analytically solvable
  2. Much better for applications, e.g., social networks
- Note:
  - Does not mean social/p2p networks are random graphs
  - But properties of social/p2p networks are well-described by random graphs

- Question:
  - How to model networks with local clusters and small diameter?
- Answer:
  - Small-world networks

Milgram's Small World Experiment



COMMUNICATIONS PROJECT

322 EMERSON HALL   HARVARD UNIVERSITY   CAMBRIDGE, MASSACHUSETTS 02138

We need your help in an unusual scientific study carried out at Harvard University. We are studying the nature of social contact in American society. Could you, as an active American, contact another American citizen regardless of his walk of life? If the name of an American citizen were picked out of a hat, could you get to know that person using only your network of friends and acquaintances? Just how open is our "open society"? To answer these questions, which are very important to our research, we ask for your help.

You will notice that this letter has come to you from a friend. He has aided this study by sending this folder on to you. He hopes that you will aid the study by forwarding this folder to someone else. The name of the person who sent you this folder is listed on the Roster at the bottom of this sheet.

In the box to the right you will find the name and address of an American citizen who has agreed to serve as the "target person" in this study. The idea of the study is to transmit this folder to the target person using only a chain of friends and acquaintances.

TARGET PERSON

Name, address, and information about the target person is placed here.

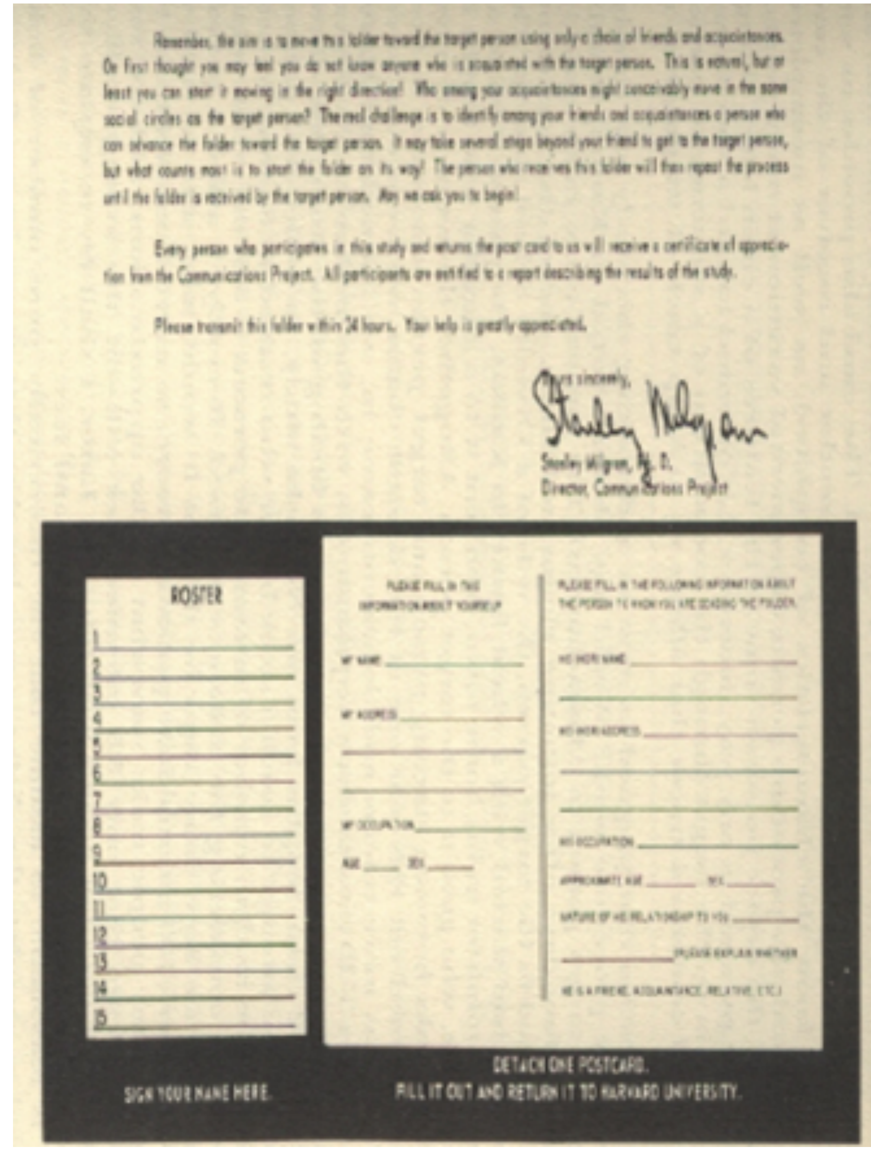
HOW TO TAKE PART IN THIS STUDY

1 ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

2 DETACH ONE POSTCARD. FILL IT OUT AND RETURN IT TO HARVARD UNIVERSITY. No stamp is needed. The postcard is very important. It allows us to keep track of the progress of the folder as it moves toward the target person.

3 IF YOU KNOW THE TARGET PERSON ON A PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIM (HER). Do this only if you have previously met the target person and know each other on a first name basis.

4 IF YOU DO NOT KNOW THE TARGET PERSON ON A PERSONAL BASIS, DO NOT TRY TO CONTACT HIM DIRECTLY. INSTEAD, MAIL THIS FOLDER (POST CARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder on to a friend, relative, or acquaintance, but it must be someone you know on a first name basis.



Remember, the aim is to move this folder toward the target person using only a chain of friends and acquaintances. On first thought you may feel you do not know anyone who is acquainted with the target person. This is natural, but at least you can start it moving in the right direction! Who among your acquaintances might conceivably move in the same social circles as the target person? The real challenge is to identify among your friends and acquaintances a person who can advance the folder toward the target person. It may take several steps beyond your friend to get to the target person, but what counts most is to start the folder on its way! The person who receives this folder will then repeat the process until the folder is received by the target person. May we ask you to begin!

Every person who participates in this study and returns the post card to us will receive a certificate of appreciation from the Communications Project. All participants are entitled to a report describing the results of the study.

Please transmit this folder within 24 hours. Your help is greatly appreciated.

Yours sincerely,

Stanley Milgram

Stanley Milgram, Ph. D.
Director, Communications Project

ROSTER

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

SIGN YOUR NAME HERE.

PLEASE FILL IN THIS INFORMATION ABOUT YOURSELF

MY NAME

MY ADDRESS

MY OCCUPATION

AGE   SEX

PLEASE FILL IN THE FOLLOWING INFORMATION ABOUT THE PERSON TO WHOM YOU ARE SENDING THE FOLDER.

HIS (HER) NAME

HIS (HER) ADDRESS

HIS OCCUPATION

APPROXIMATE AGE   SEX

NATURE OF HIS RELATIONSHIP TO YOU

(PLEASE EXPLAIN WHETHER HE IS A FRIEND, ACQUAINTANCE, RELATIVE, ETC.)

DETACH ONE POSTCARD.
FILL IT OUT AND RETURN IT TO HARVARD UNIVERSITY.

Six Degrees of Separation

- Famous experiment from 1960's (S. Milgram)
- Send a letter to random people in Kansas and Nebraska and ask people to forward letter to a person in Boston (~ 3000 km)
  - Person identified by name, profession, and city
- Rule: Give letter only to people you know by first name and ask them to pass it on according to same rule
  - Some letters reached their goal
- Letter needed six steps on average to reach the person
- Graph theoretically: Social networks have dense local structure, but (apparently) small diameter
  - Generally referred to as "small world effect"
  - Usually, small number of persons act as "hubs"

Small-World Networks

- Discovered by Watts and Strogatz (1998) (30 years after Milgram)
- Watts and Strogatz looked at three networks
  - Film collaboration between actors, US power grid, Neural network of worm Caenorhabditis elegans ("C. elegans")
- Measured characteristics
  - CC as a measure for 'regularity', or 'locality' of the network
    - If it is high, edges are rather build between close nodes
  - CPL - characteristic path length
- Results
  - Grid-like networks
    - <u>High</u> CC & <u>high</u> CPL (edges are not 'random', but rather 'local')
  - Most real-world (natural) networks
    - <u>High</u> CC (0.3-0.4) & <u>low</u> CPL

Small-World Networks and Random Graphs

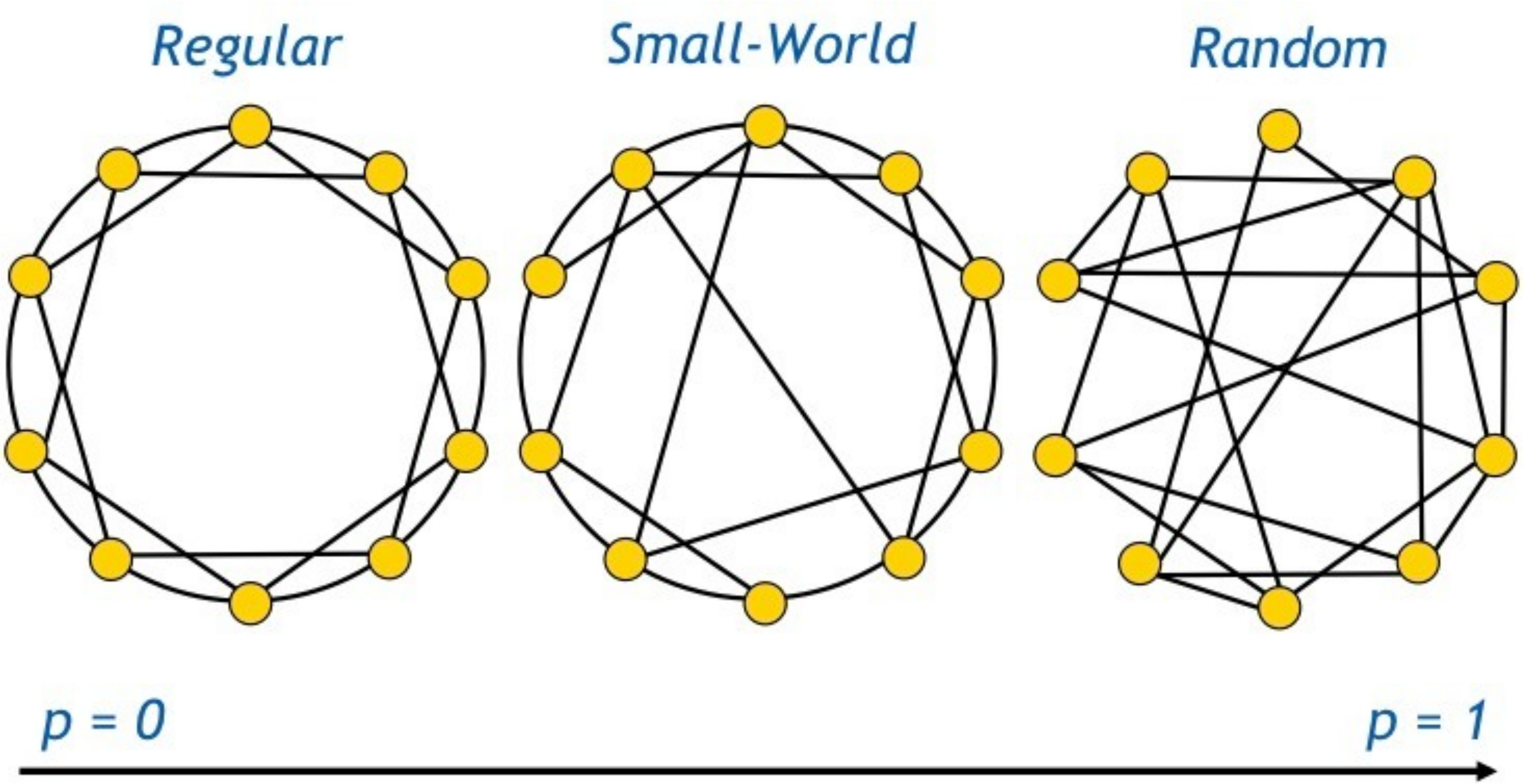
- Results
  - Compared to a random graph with same number of nodes
  - Diameters similar, slightly higher for real graph
  - Clustering coefficient orders of magnitude higher
- Definition of small-worlds network
  1. Dense local clustering structure
  2. Small diameter comparable to that of a same-sized random graph

	CPL (actual)	CPL (random)	CC (actual)	CC (random)
Film collaboration	3.65	2.99	0.79	0.00027
Power Grid	18.7	12.4	0.08	0. 005
C. elegans	2.65	2.25	0.28	0.05

The WS Model - Constructing Small-World Graphs

- Put all n nodes on a ring, number them consecutively from 1 to n
  - Connect each node with its k clockwise neighbors
  - Traverse ring in clockwise order
- For every edge
  - Draw random number r
  - If $r < p$
    - Re-wire edge to random target node (no duplicates)
  - Else
    - Keep “old” edge
- Different values of p give different graphs
  - If p is close to 0, then original structure mostly preserved
  - If p is close to 1, then new graph is random
  - Interesting things happen when p is somewhere in-between

Regular, Small-World, Random

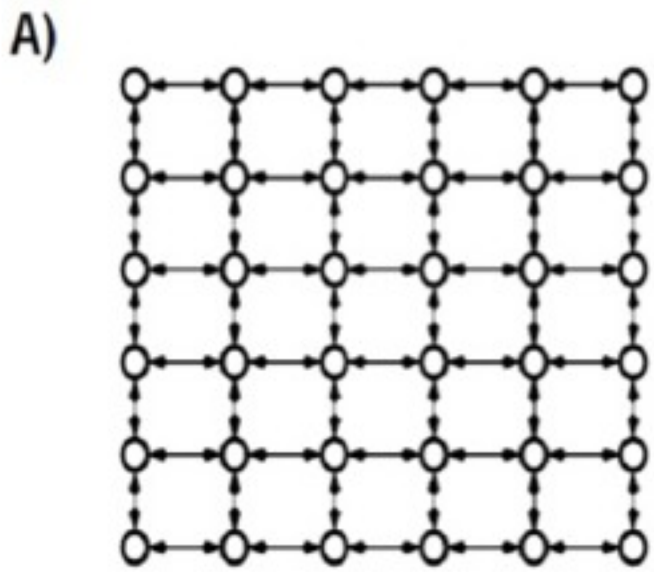


Kleinberg's Small-World Navigability Model

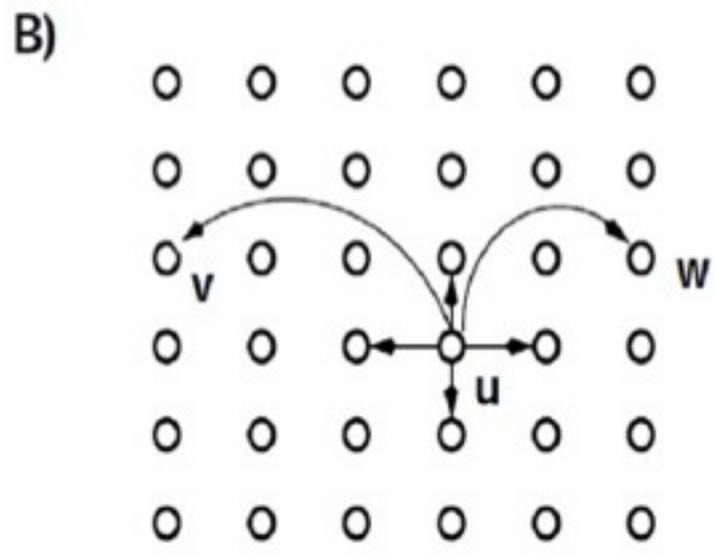
- Small-world model and power law explain why short paths exist
- Missing piece in the puzzle: why can we find these paths?
- Setting
  - Each node has only local information
  - Even if a short cut exists, how do people know about it?
  - Milgram's experiment:
    - Some additional information (profession, address, hobbies etc.) is used to decide which neighbor is "closest" to recipient
    - Results showed that first steps were the largest
- Kleinberg's Small-World Model (2d)
  - Set of points in an n x n grid
  - Distance is the number of "steps" separating points
  - $d(i, j) = |x_i - x_j| + |y_i - y_j|$ (Manhattan distance)

Kleinberg's Small-World Navigability Model (2d)

- Connections on the grid
  - All nodes are connected to all nodes at distance p
- Long-range contacts to nodes at larger distances (q)
  - With probability decreasing with growing distance



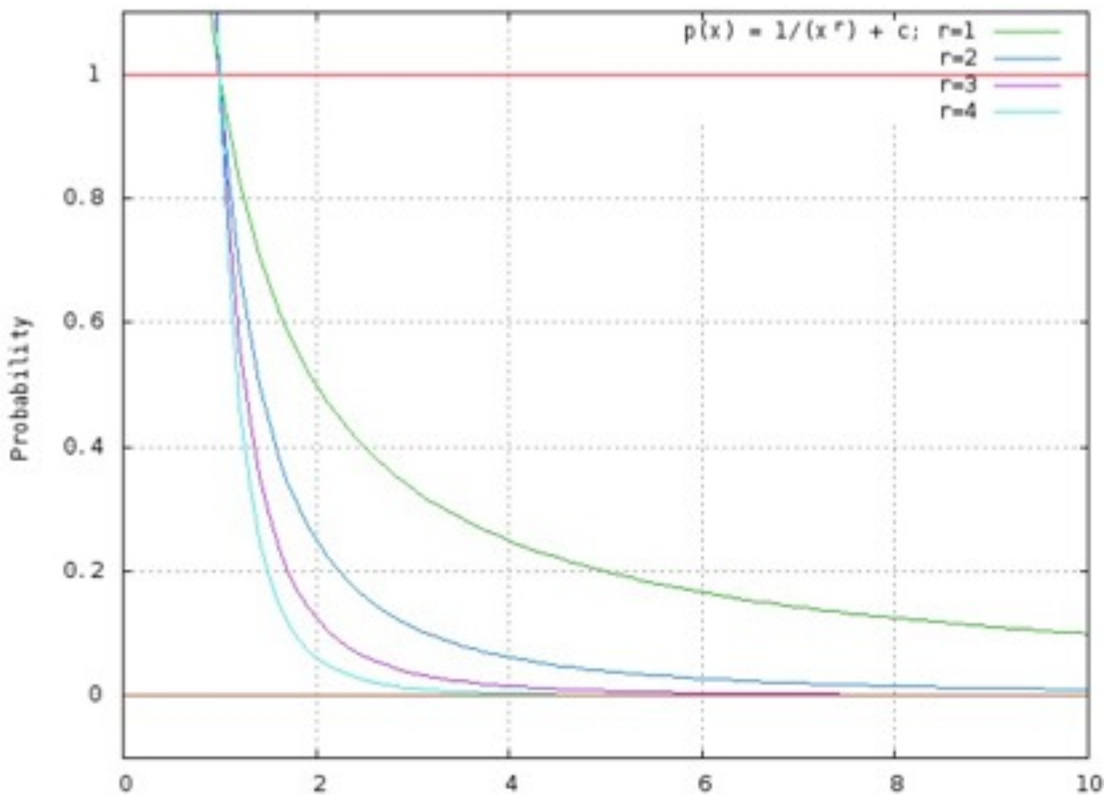
p = 1, q = 0



p = 1, q = 2

Kleinberg's Small-World Navigability Model

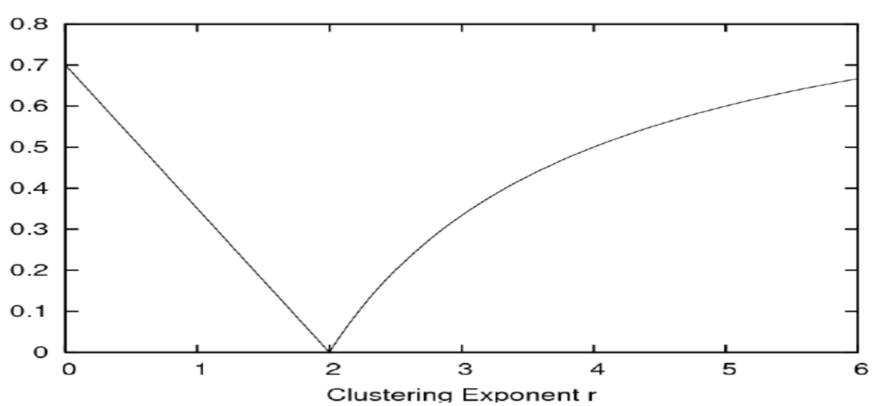
- d-dimensional identifier space
  - Every node v_i is connected to node v_j within distance p
  - For every node v_i, additional q long-range edges are added
  - Probability that node v_j is selected is proportional to $d(v_i, v_j)^{-r}$
    - For constant r (harmonic distribution p)



$$p_{sw}(v_i) = \frac{1}{x^r} + c$$

Navigation in Kleinberg's Model

- Simple greedy routing (2d)
  - Nodes only know local links and target position
  - Always use the link that brings message closest to target
    - If r=2 (p=q=1), expected lookup time is $O(\log^2 n)$
    - If r≠2, expected lookup time is O(nε), where ε depends on r
- Kleinberg has shown
  - Routing takes $O(\log^2 n)$ hops iff r=d (d = number of dimensions)
    - Idea behind proof
      - For any r < s there are too few random edges to make paths short
      - For r > s there are too many random edges / too many choices
      - The message will make a (long) random walk through the network



Summary of Kleinberg's Model

- Kleinberg's small world model thus provides
  - Way of building a peer-to-peer overlay network
  - Very simple, greedy, and local routing protocol is applicable

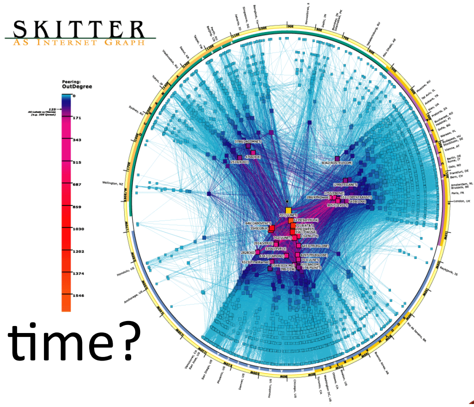
- Routing: forward message to contact who is closest to target
  - Assumes some way of associating nodes with points in grid
  - Assumes some way to know about "closest" ones
  - Compare with CAN

Problems with Small-World Graphs

- Small-world graphs explain why:
  - Highly clustered graphs can have short average path lengths (“short cuts”)
- Small-world graphs do NOT explain why:
  - This property emerges in real networks
  - Real networks are practically never ring- or grid-like
- Further problem with small-world graphs:
  - Nearly all nodes have same degree
  - Not true for random graphs
  - What about real networks?

Internet

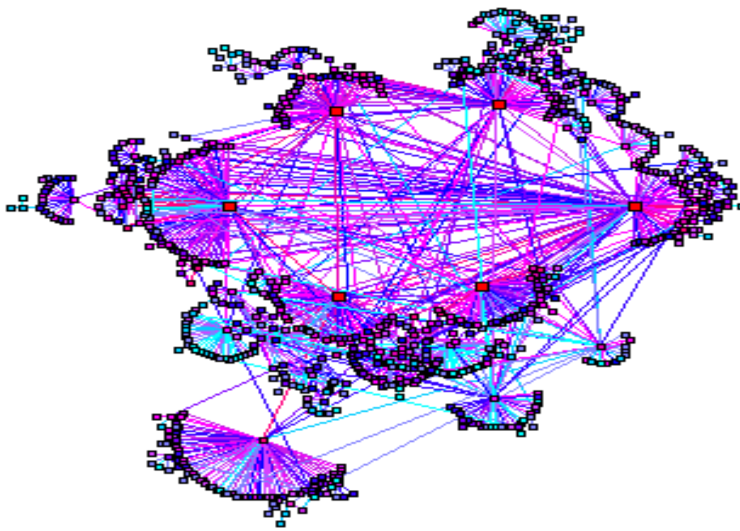
- Faloutsos et al. study from 99: Internet topology examined in 1998
- AS-level topology, during 1998 Internet grew 45%

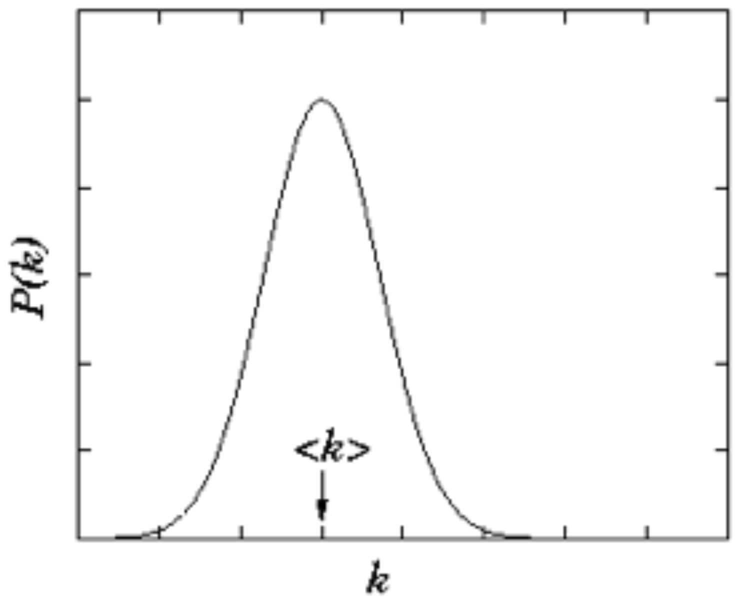


- Motivation:
  - What does the Internet look like?
  - Are there any topological properties that don't change over time?
  - How to generate Internet-like graphs for simulations?
- 4 key properties found, each follows a power-law; Sort nodes according to their (out)degree
  - Outdegree of a node is proportional to its rank to the power of a constant
  - Number of nodes with same out-degree is proportional to the out-degree to the power of a constant
  - Eigenvalues of a graph are proportional to the order to the power of a constant
  - Total number of pairs of nodes within a distance d is proportional to d to the power of a constant

World Wide Web

- Links between documents in the World Wide Web
  - 800 Mio. documents investigated (S. Lawrence, 1999)
- What was expected so far?
  - Number of links per web page: k ~ 6
  - Number of pages in the WWW: $N_{WWW} \sim 10^9$



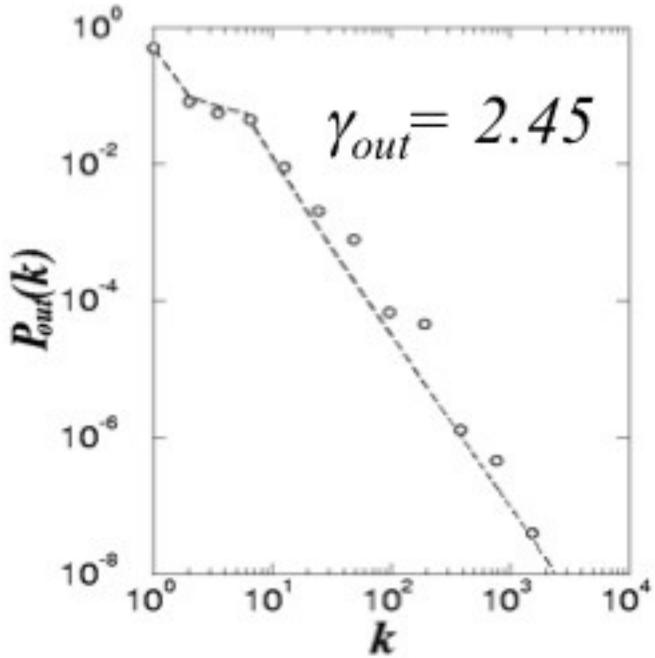


Probability “page has 500 links”.
$P(k=500) \sim 10^{-99}$

Number of pages with 500 links:
$N(k=500) \sim 10^{-90}$

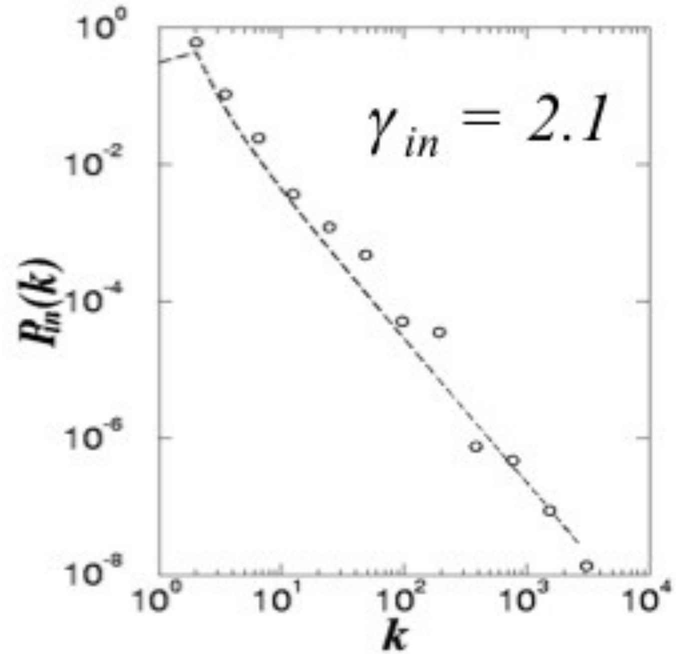
World Wide Web: Result of Investigation

P(page has k links)



$$P_{out}(k) \sim k^{-\gamma\, out}$$

P(k pages link to this page)



$$P_{in}(k) \sim k^{-\gamma\, in}$$

Probability "page has 500 links".

~~$P(k=500) \sim 10^{-99}$~~

$P(k=500) \sim 10^{-6}$

Number of pages with 500 links:

~~$N(k=500) \sim 10^{-90}$~~

$N(k=500) \sim 10^{3}$

Power Law Networks

- Also known as scale-free networks
- "Power Law" relationship for Web pages
  - The probability P(k) that a page has k links (or k other pages link to this page) is proportional to the number of links k to the power of y
  - $P(k) \sim ck^{-y}$
- General "Power Law" Relationships
  - A certain characteristic k is – independent of the growth of the system – always proportional to k^a, whereby a is a constant (often -2 < a < -4)
- Power laws very common ("natural")
  - And power law networks exhibit small-world-effect
  - E.g. WWW: 19 degrees of separation
  - (R. Albert et al, Nature (99); S. Lawrence et al, Nature (99))

Examples for Power Law Networks

- Economics
  - Pareto: income distribution
    (common simplification: 20% of population own 80% of the wealth)
  - Sizes of companies and cities (Zipf's law)
- Human networks
  - Professional (e.g. collaborations between actors, scientists)
  - Social (friendship, acquaintances)
  - Sexual-contact networks
- Many other natural occurrences
  - Distribution of English words (Zipf's law again)
  - Areas burnt in forest fires, meteor impacts on the moon
- Internet also follows some power laws
  - Popularity of Web pages (possibly related to Zipf's law for English words?)
  - Connectivity of routers and Autonomous Systems, Gnutella's topology!

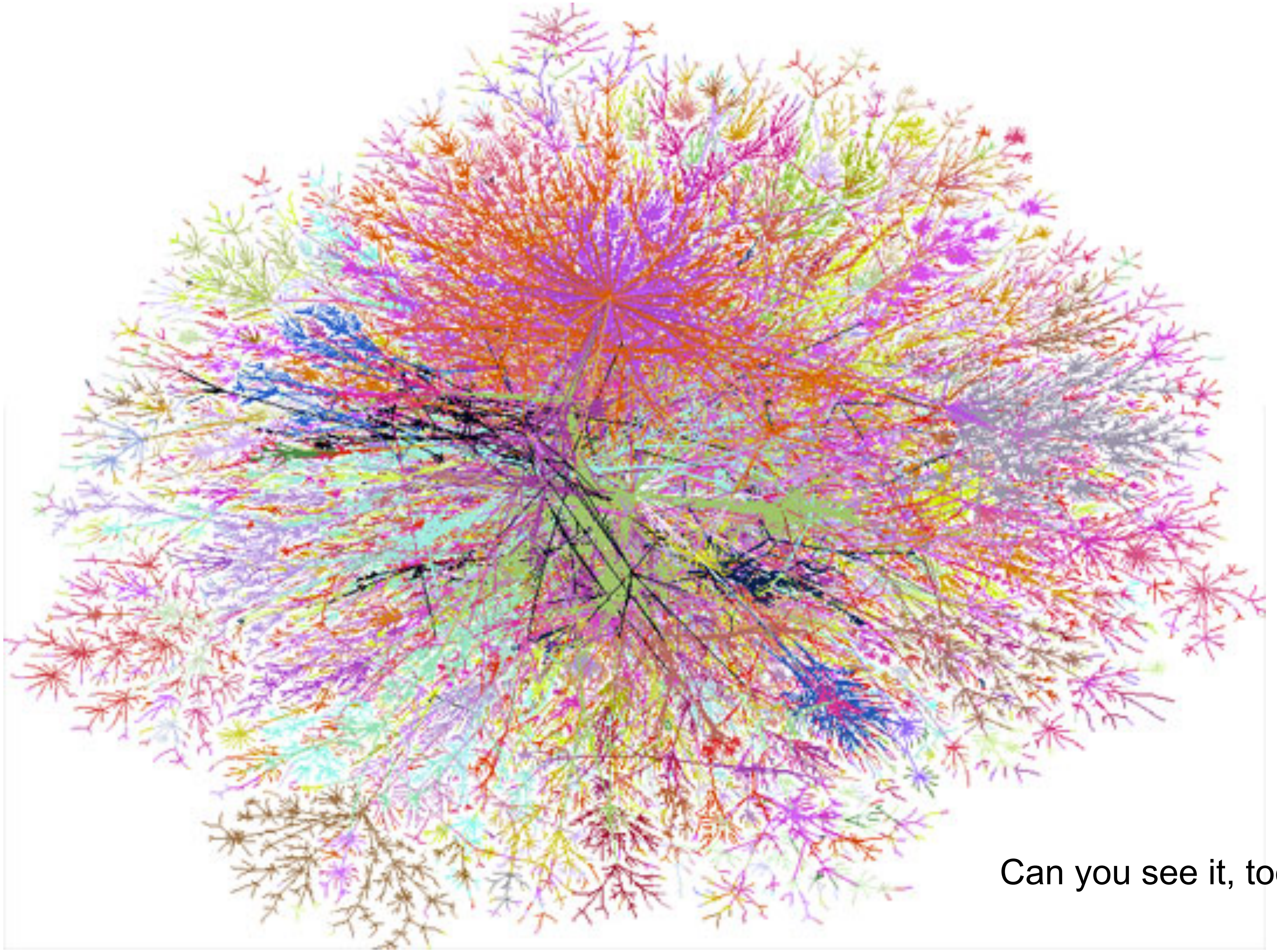
Graph Generation / Analysis: So What!?

- Science is the art of systematically trying to understand nature
- Today more than ever: yes, we can!
- In our case:
  - Understand the properties of different networks / topologies / models
  - Define which of them are „good“ and „bad“ (wrt. certain requirements)
  - Understand topology classes (are there short-cuts we haven‘t even seen?)
  - Looking for ways to optimize topologies
  - Re-thinking previous designs
    - IP and routing are optimized for random graphs
  - Require topology generators
    - Allows us to test our algorithms in them (the freenet case)

Barabasi-Albert Model

- How do power law networks emerge?
  - In a network where new nodes are added and new nodes tend to connect to well-connected nodes, the vertex connectivities follow a power-law
- Barabasi-Albert Model
  - Power-law network is constructed with two rules:
  1. Network grows in time
  2. New node has preferences to whom it wants to connect
- Preferential connectivity modeled as
  - Each new node connects to m other nodes
  - Probability of connecting to v_j is proportional to its degree $d(v_j)$
- New nodes tend to connect to well-connected nodes
  - Another way of saying this: “The rich get richer”

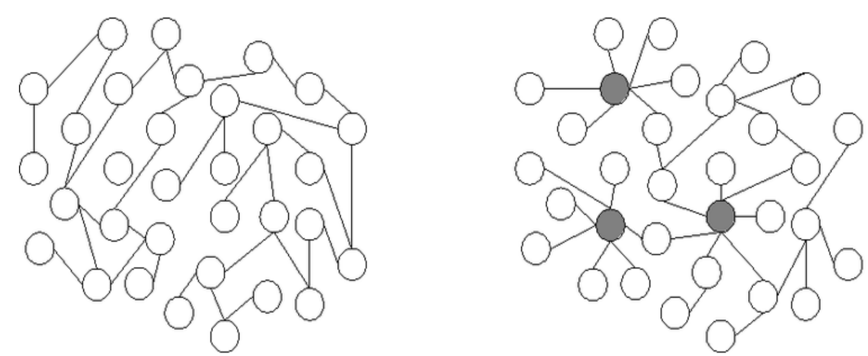
Barabasi-Albert Model



Can you see it, too? ;-)

Copying model

- Alternative generative model (R. Kumar, P. Raghavan, et al. 2000)
  - In each step: randomly copy one of the existing nodes keeping all its links
    - Connect original node and copy
  - Randomly remove edges from both nodes with a very small probability
    - For each removed edge randomly draw new target nodes
- Probability of node v getting a new edge in some time step is proportional to its degree at that time
  - More edges → higher probability of a neighbor being chosen during a step
  - Clear contrast to random networks
    - Small number of well-connected hubs
    - Many nodes with few connections



(a) Random network (b) Scale-free network

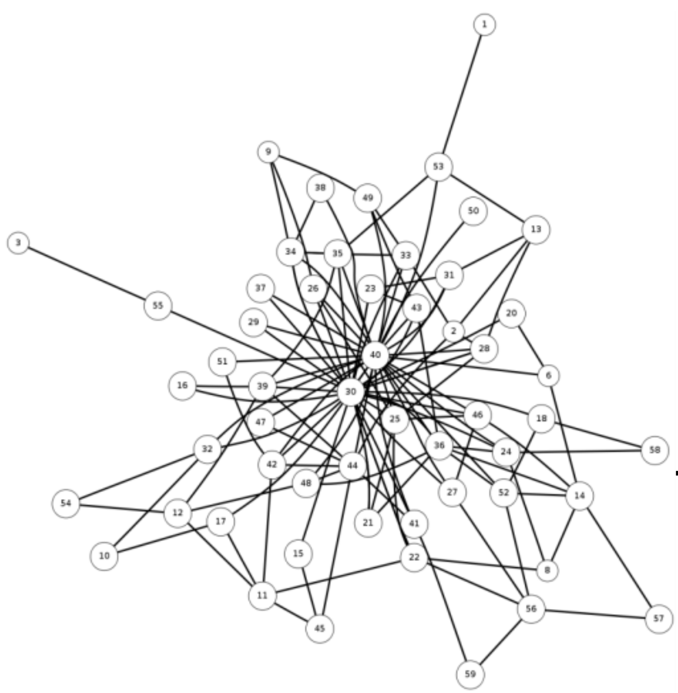
Further Need for Improvement of Generators

- Social Graphs, P2P topologies, or generated AS-level (and general) Internet topologies in many cases still not very realistic

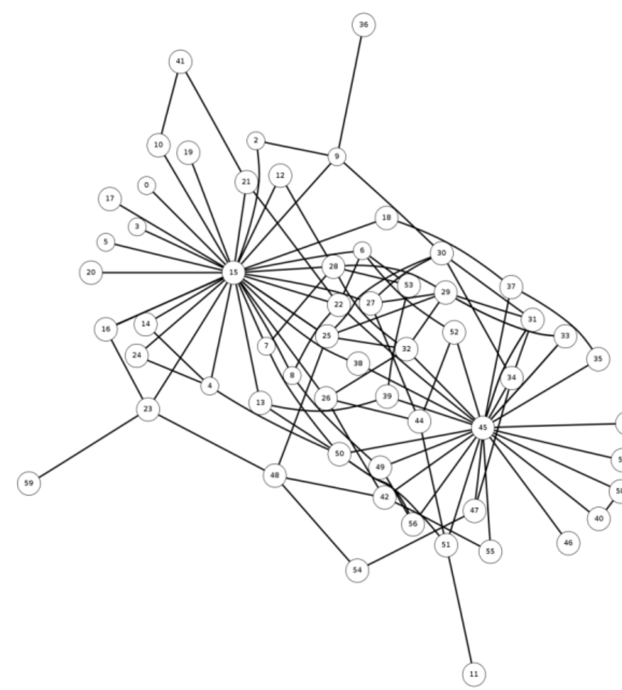
- All follow power-law, but which could be modeling the Internet?

Further Metrics: Dis-/Assortativity

- Who is interconnected?
- How „similar“ are the connected nodes?
- Assortativity refers to the preference of nodes to connect to other nodes that in some way are similar.
- In Internet/Overlay/Social Network analysis the „way“ usually refers to the node degrees:



- the Internet?

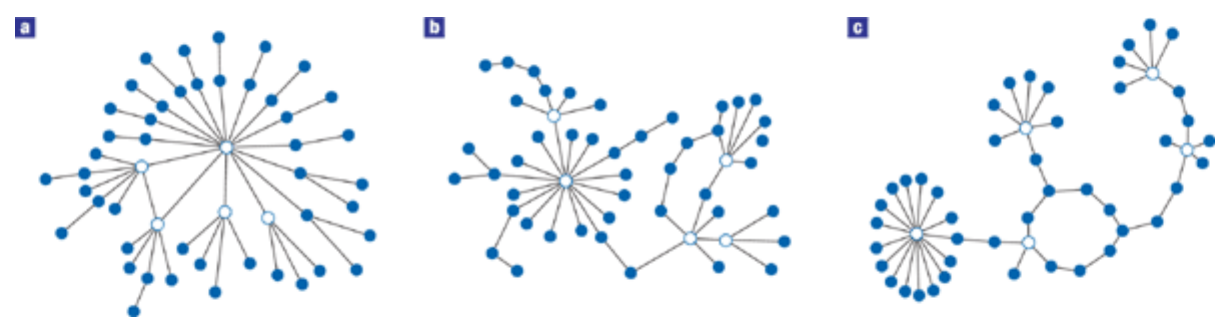


Assortativity in Real Networks

- Social Networks exhibit clear assortativity
- Technological/biological networks exhibit clear disassortativity

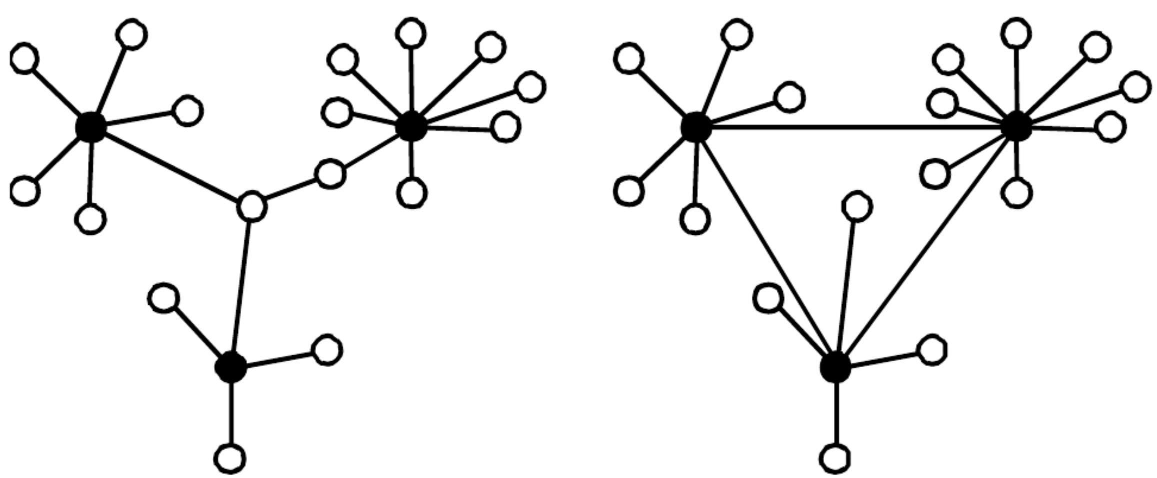
Network	n	r
Physics coauthorship (a)	52 909	0.363
Biology coauthorship (a)	1 520 251	0.127
Mathematics coauthorship (b)	253 339	0.120
Film actor collaborations (c)	449 913	0.208
Company directors (d)	7 673	0.276
Internet (e)	10 697	−0.189
World-Wide Web (f)	269 504	−0.065
Protein interactions (g)	2 115	−0.156
Neural network (h)	307	−0.163
Marine food web (i)	134	−0.247
Freshwater food web (j)	92	−0.276
Random graph (u)		0
Callaway *et al.* (v)		$\delta/(1+2\delta)$
Barabási and Albert (w)		0

- Again: which one could be the Internet? ;-)



Is it really? Power-Law and Disassortative?

- What about peering points?
- Peering points generally have a very high node degree and are connected to each other
- Introducing: The Rich-Club Connectivity
  - Measures, which fraction of nodes with the highest node degree are actually interconnected



(a) (b)

Fig 5 Two scale-free networks — (a) without a rich-club, and (b) with a rich-club.

The Internet and the Rich Club Connectivity

- Contradiction?!?
  – The Internet is disassortative and Rich-Club Connected?
- Only partly
  – The low- and mid-degree nodes are disassortative, the rest is the „rich club“

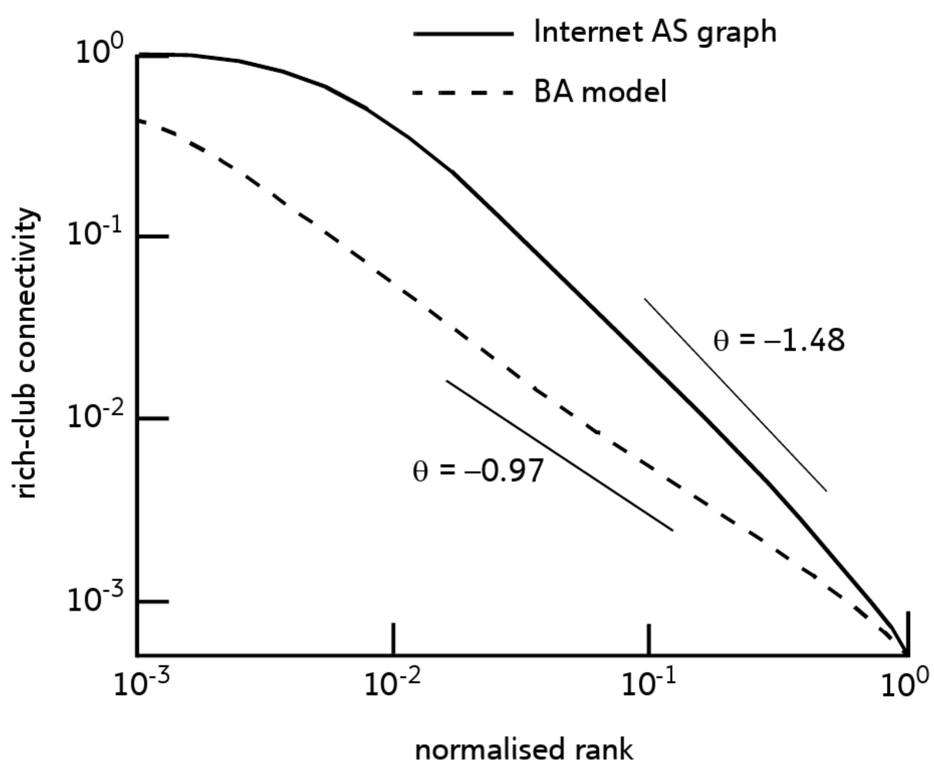
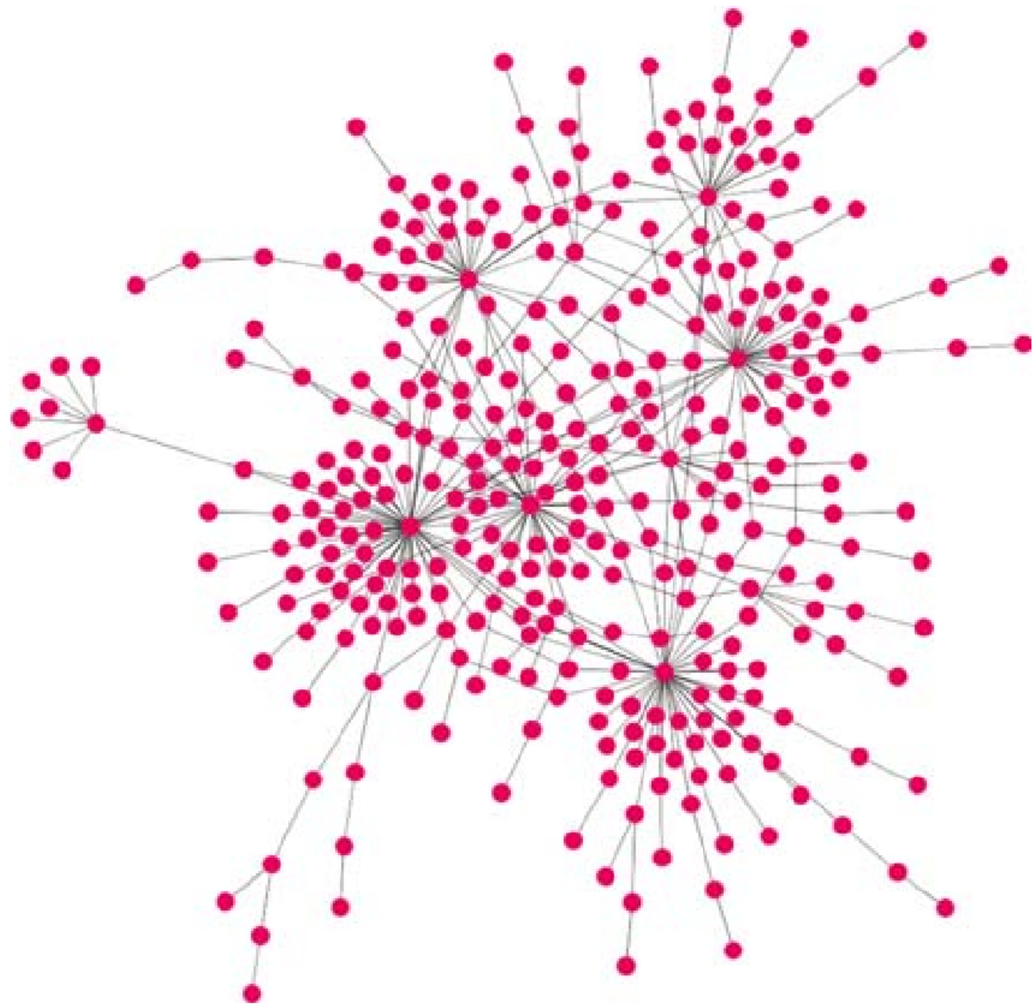


Fig 6 Rich-club connectivity versus normalised rank, $\phi(r/N)$.

Great! We Can Generate Internet-like Topologies!

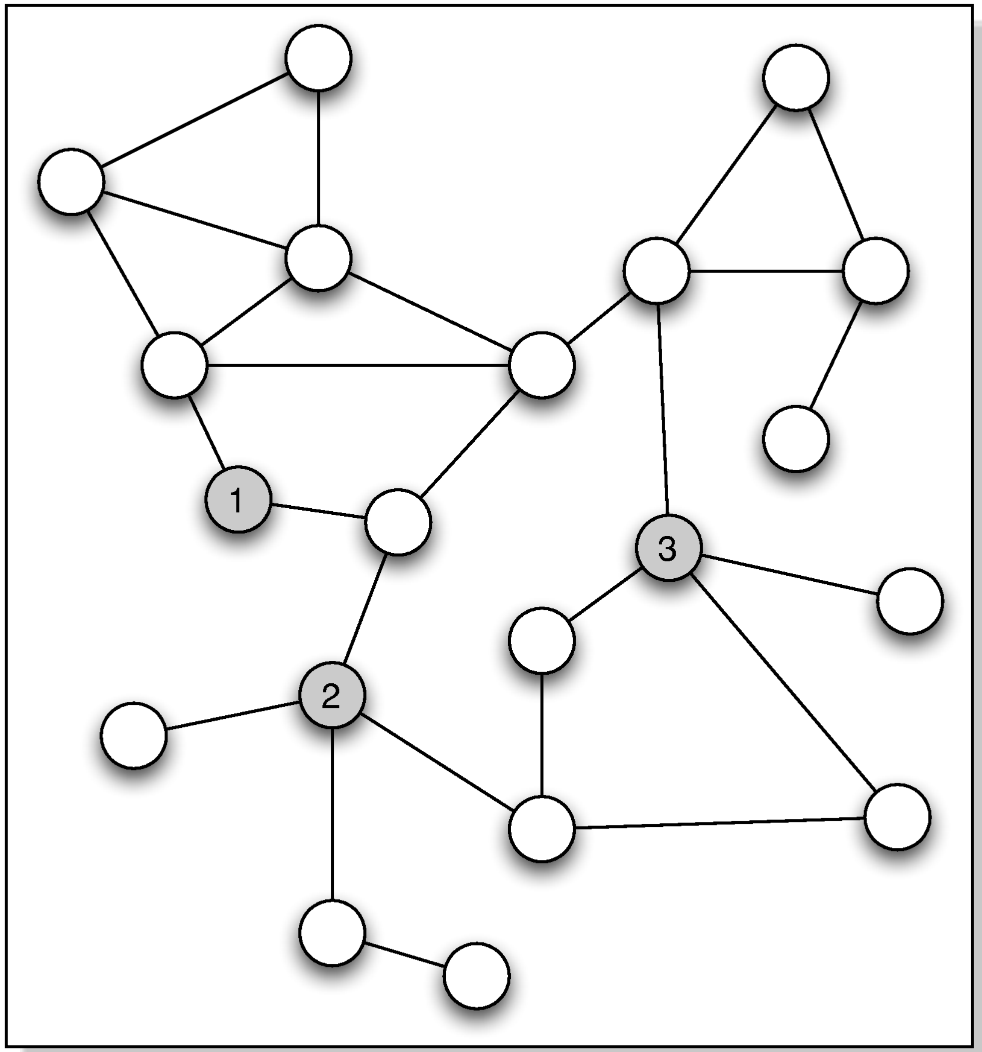
- Useful if we want to evaluate a distributed system
- Make sure the properties of the underlay are correctly assumed

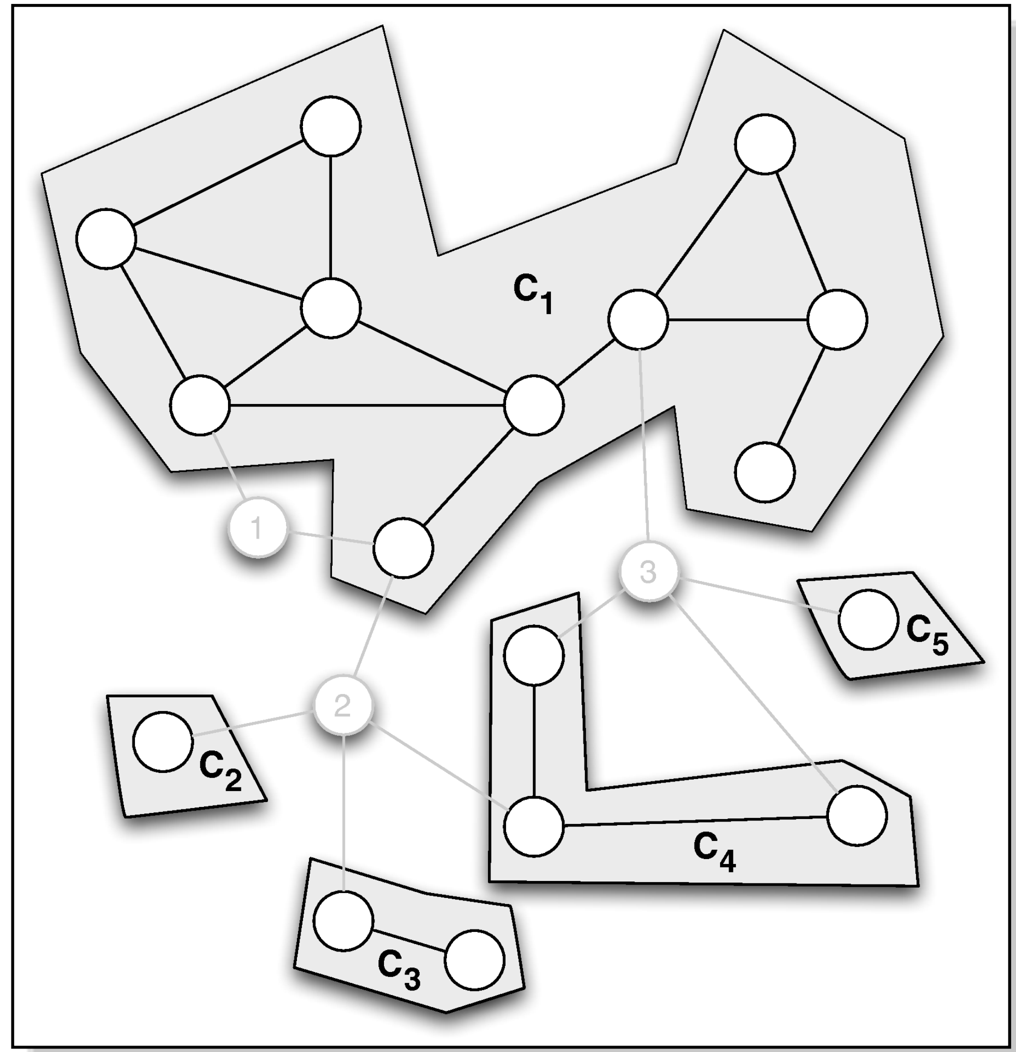


Properties of Scale-Free Networks

- Recall: „Easy“ to find paths in small world networks
- How about robustness?
  - Further robustness/resistance metrics:
  - Balanced (vertex) cut
    - Minimum number of edges (vertices) removed to achieve a network fragmentation into two equal components
  - Average Connected Distance
    - CPL under failure of or attack on nodes
  - Maximum Isolated Component Size (giant connected component)
  - Average Isolated Component Size
  - „Point of Rupture“
    - Minimum number of nodes the removal of which causes the network to fragment into components, with $|GCC| < 0.5*|V|$

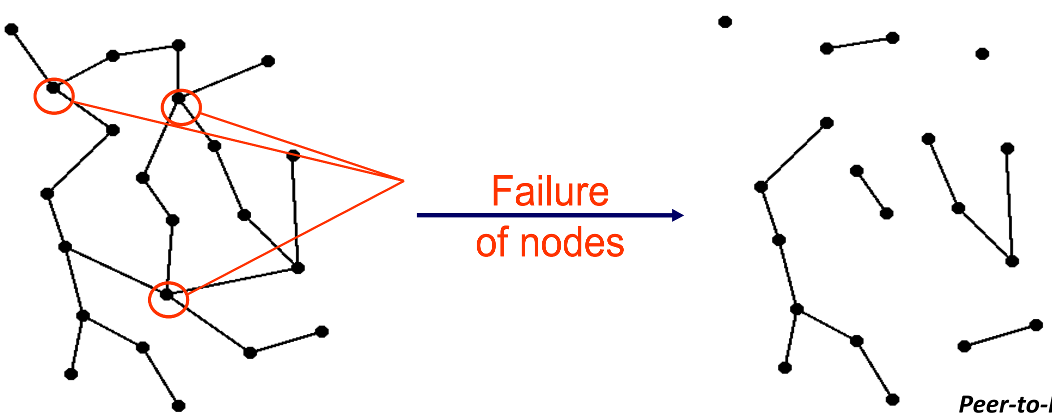
Robustness: Example





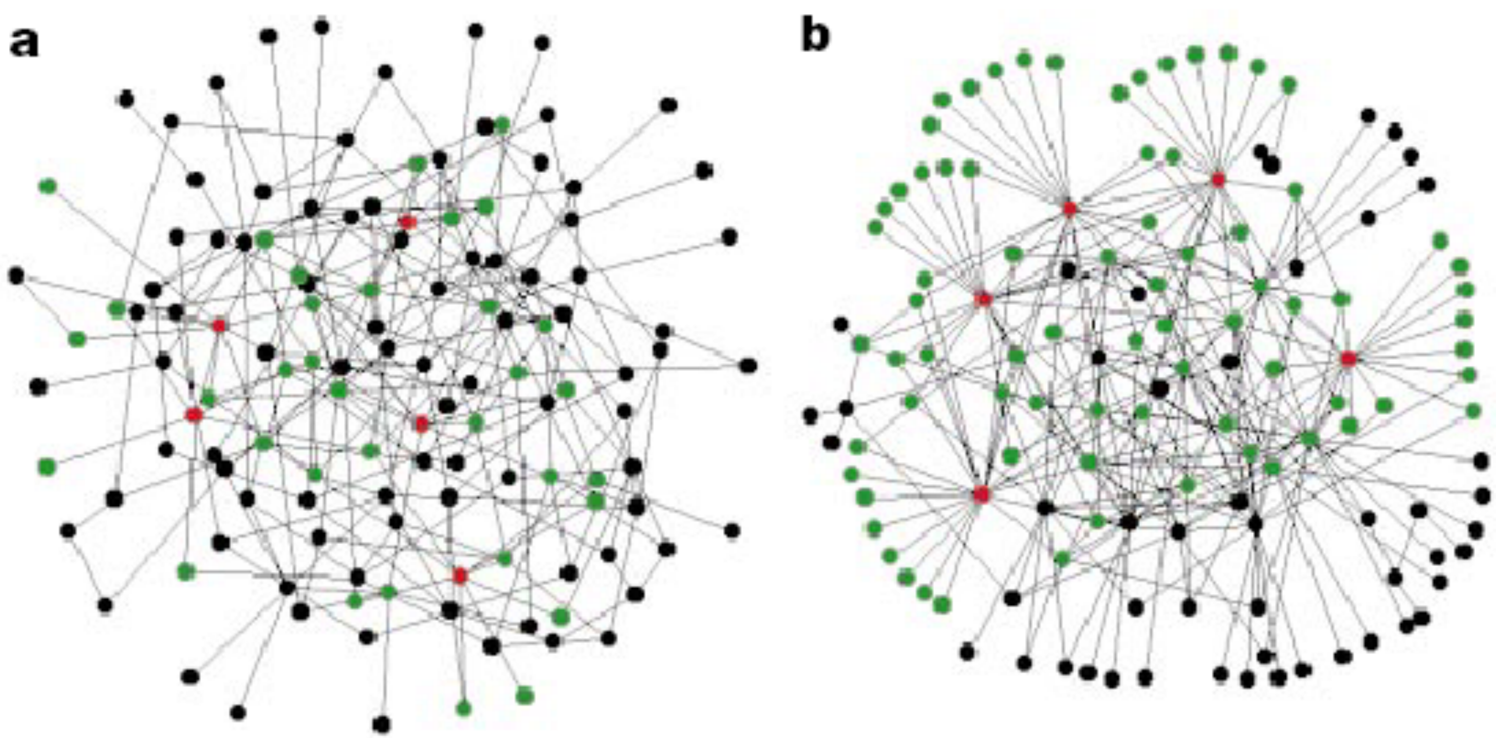
Properties of Scale-Free Networks

- Robustness against random failures
  - Important property of networks with scale-free degree distribution
  - Remove randomly chosen vertex v
    - With high probability => damage to the network small
- But very sensitive against attacks
  - Adversary removes highest degree vertices first
  - The network quickly decomposes into very small components
- Note: random graphs are not robust against random failures
  - But not sensitive against attacks either
  - Because all vertices more or less have the same degree



Robustness of Scale Free vs. Random Networks

- Experiment: take network of 10k nodes (random and power-law)
- Remove nodes randomly



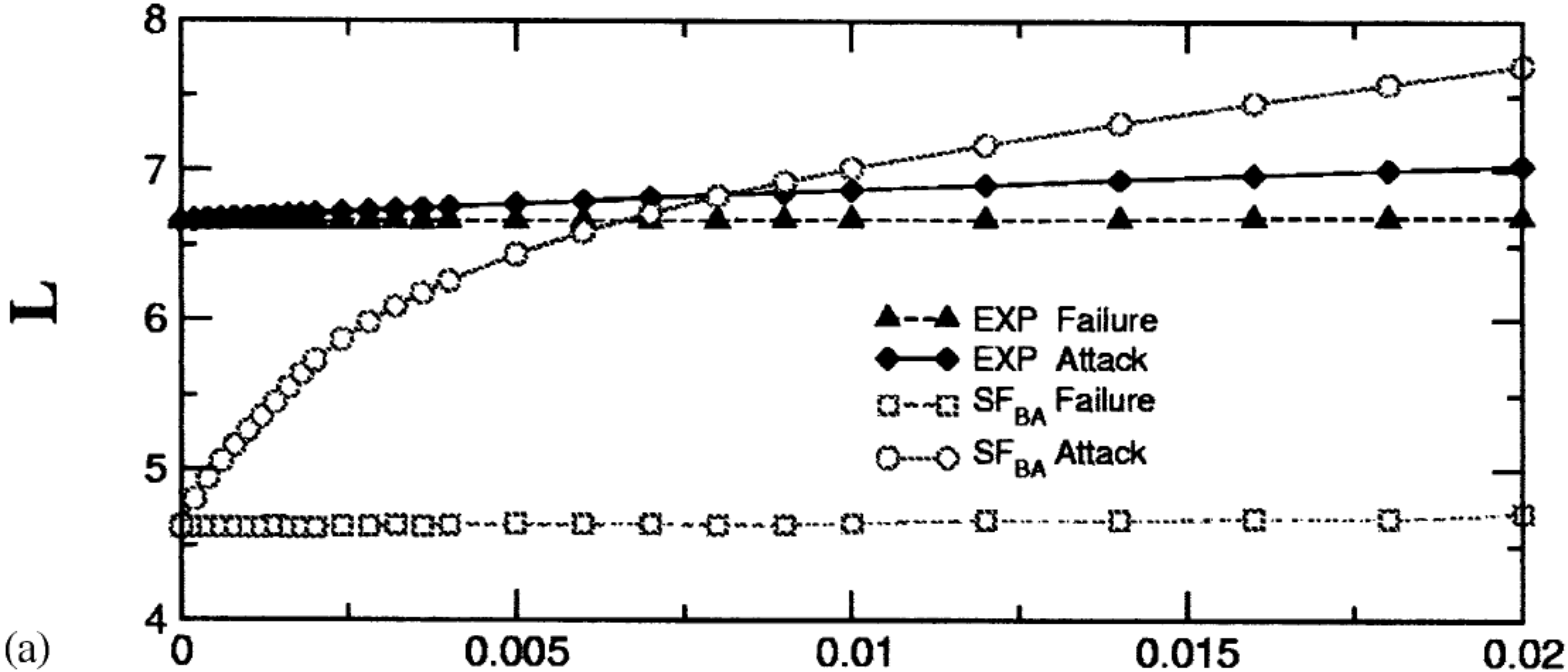
Random Graph *„Power Law“ Graph*

Robustness of Scale Free vs. Random Networks

- Random graph:
  - Take out 5% of nodes: Biggest component 9000 nodes
  - Take out 18% of nodes: No biggest component, all components between 1 and 100 nodes
  - Take out 45% of nodes: Only groups of 1 or 2 survive
- Power-law graph:
  - Take out 5% of nodes: Only isolated nodes break off
  - Take out 18% of nodes: Biggest component 8000 nodes
  - Take out 45% of nodes: Large cluster persists, fragments small
- Networks with power law exponent < 3 are very robust against random node failures

- ONLY true for random failures!

The consequence…



Unstructured P2P Networks

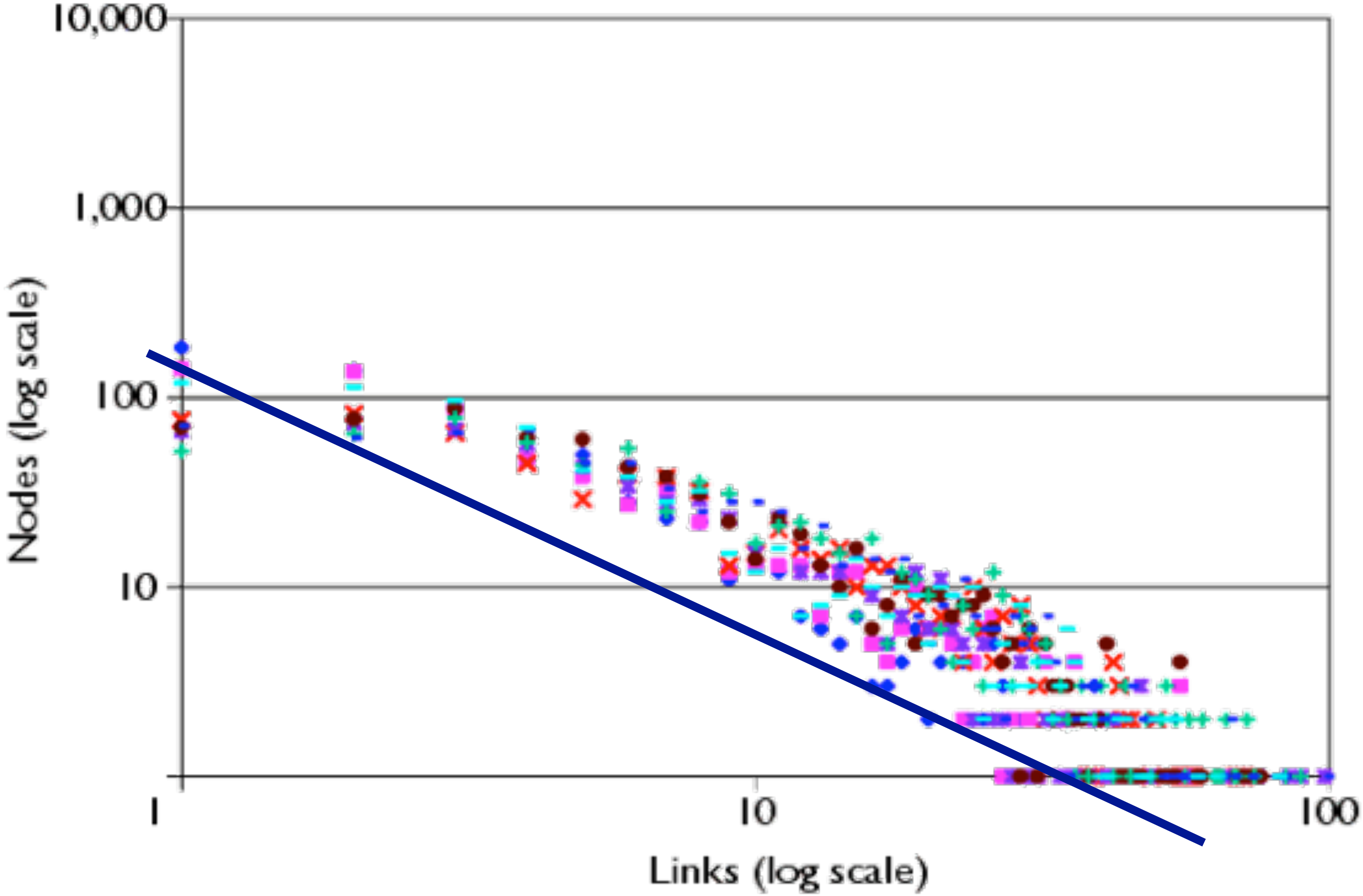
- What do real (unstructured) Peer-to-Peer Networks look like?

- Depends on the protocols used
  - Some P2P networks, e.g., Freenet, evolve voluntarily in a small-world
    - High clustering coefficient, small diameter
  - Some protocols, e.g., Gnutella, implicitly generate a scale-free degree distr.

- Case study: Gnutella network
- How does the Gnutella network evolve?
  - Nodes with high degree answer more likely to Ping messages
  - Thus, they are more likely chosen as neighbor
  - Host caches always/often provide addresses of well connected nodes

Gnutella

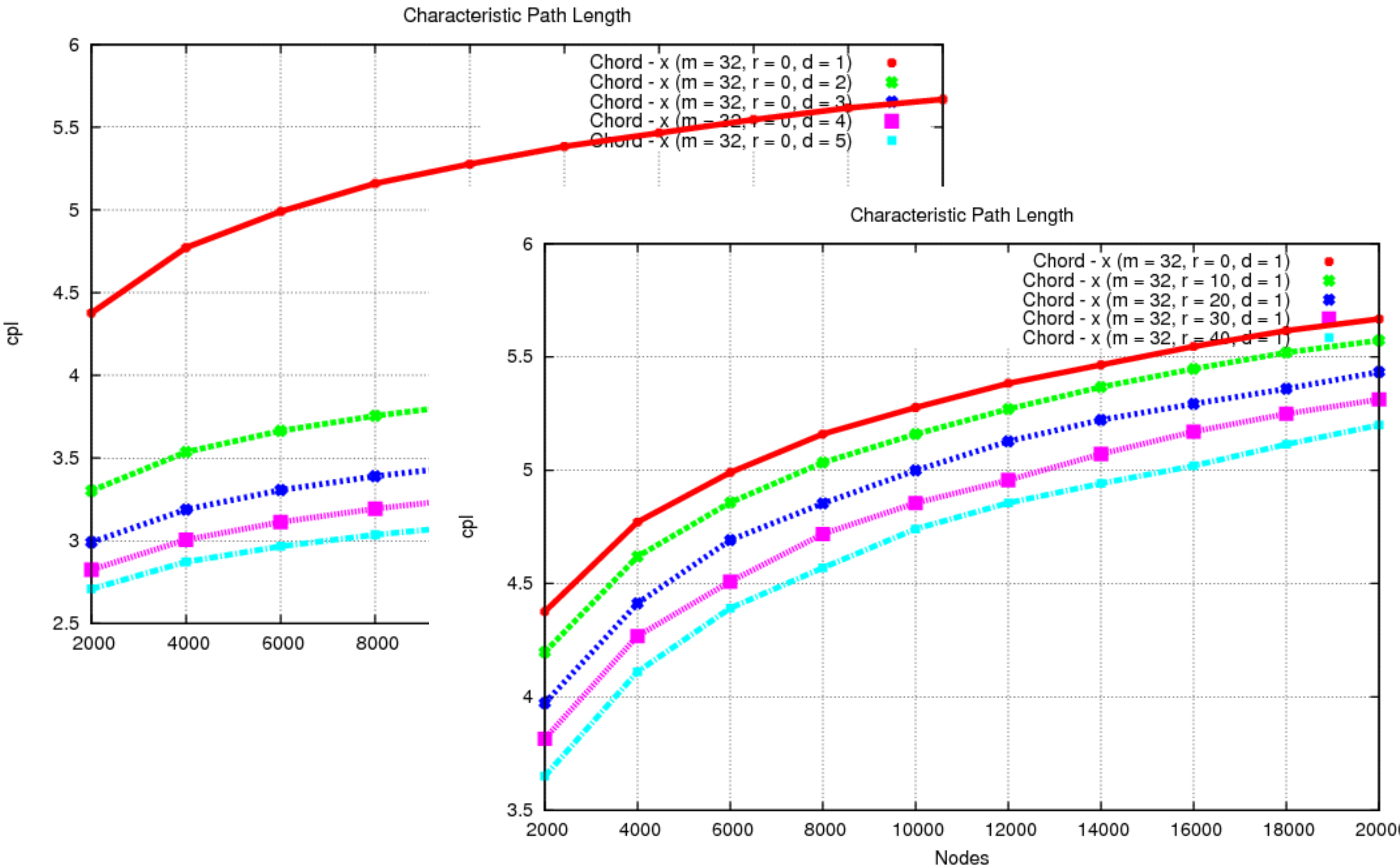
- Node degrees in Gnutella follow power-law rule



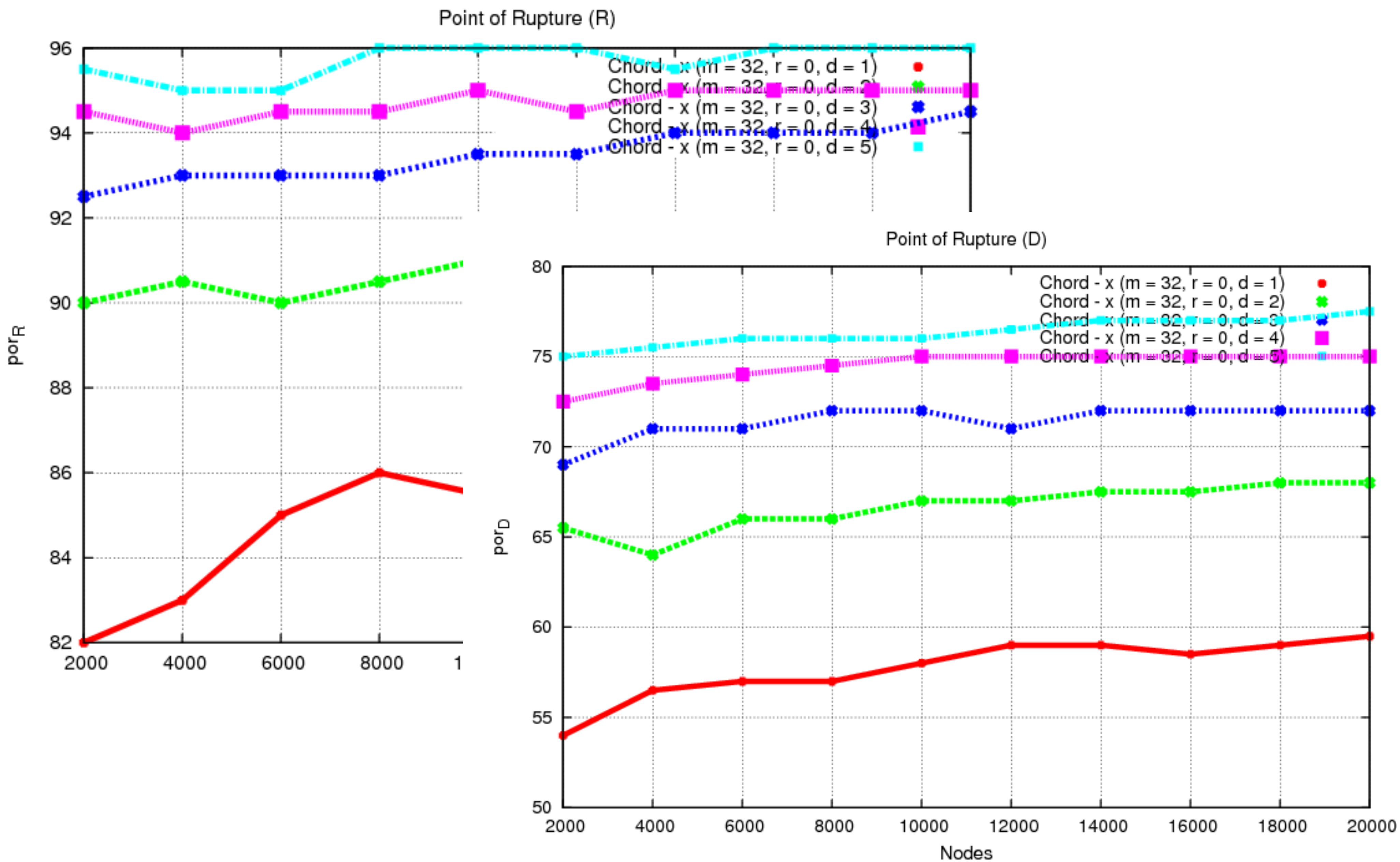
Gnutella /2

- Network diameter stayed nearly constant, though the network grew by one order of magnitude
- Robustness
  - Remember: networks with power-law exponent < 3 are very robust against random node failures (Gnutella's exponent is 2.3)
- Theoretical experiment
  - Subset of Gnutella with 1771 nodes
  - Take out random 30% of nodes, network still survives
  - Take out 4% of best connected nodes, network splinters
- For more information on Gnutella, see:
  - Matei Ripeanu, Adriana Iamnitchi, Ian Foster: Mapping the Gnutella Network, IEEE Internet Computing, Jan/Feb 2002
  - Zeinalipour-Yazti, Folias, Faloutsos, "A Quantitative Analysis of the Gnutella Network Traffic", Tech. Rep. May 2002

Analyzing Chord - Characteristic Path Length



Analyzing Chord - Point of Rupture



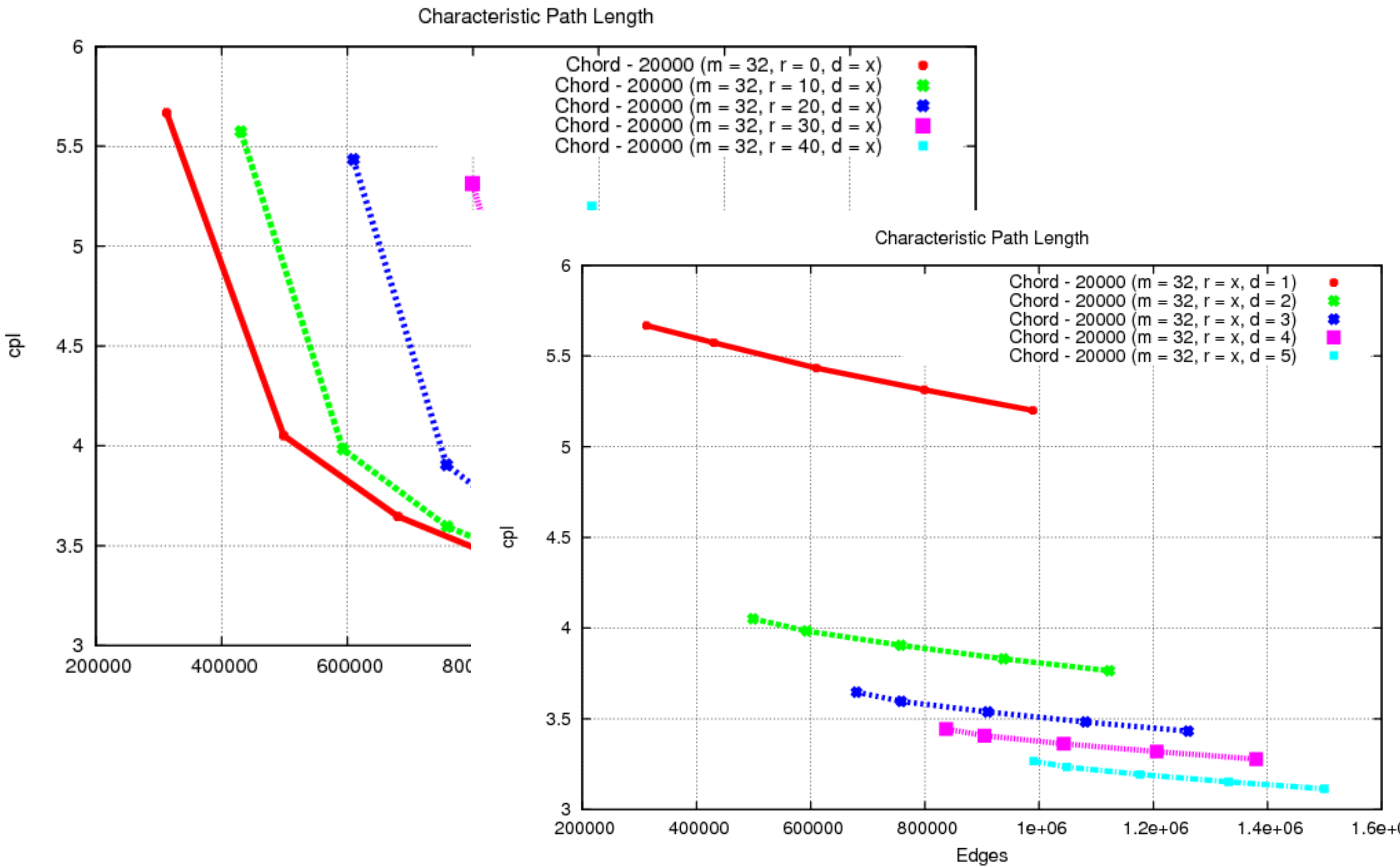
Analyzing Chord: So what can we do?

- Increase connectivity!
- Select more neighbors! But how?
  - Additional fingers: Finger[i]=n.id+(1+1/d)^i
  - Additional successors (r)
  - What will the results look like?

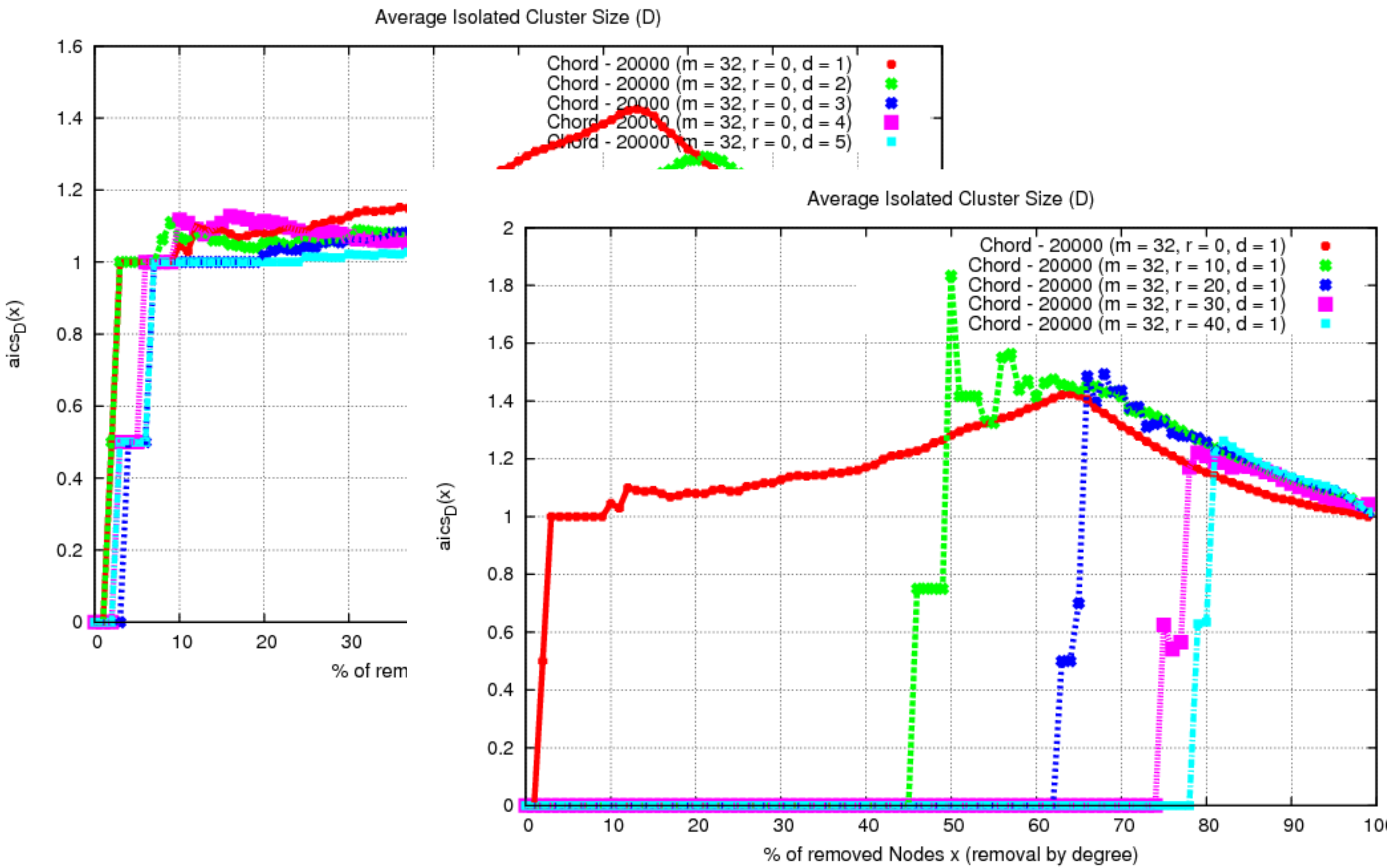
distance d	fingers$_{max}$	finger distances
d	$log_{1+1/d}\ id_{max}$	$(1+1/d)^i, 1 \leq i \leq fingers_{max}$
1	9	2, 4, 8, 16, 32, 64, 128, 256, 512
2	17	1, 2, 3, 5, 7, 11, 17, 25, 38, 57, 86, 129, 194, 291, 437, 656, 985
3	24	1, 1, 2, 3, 4, 5, 7, 9, 13, 17, 23, 31, 42, 56, 74, 99, 133, 177, 236, 315, 420, 560, 747, 996
4	31	1, 1, 1, 2, 3, 3, 4, 5, 7, 9, 11, 14, 18, 22, 28, 35, 44, 55, 69, 86, 108, 135, 169, 211, 264, 330, 413, 516, 646, 807, 1009
5	38	1, 1, 1, 2, 2, 2, 3, 4, 5, 6, 7, 8, 10, 12, 15, 18, 22, 26, 31, 38, 46, 55, 66, 79, 95, 114, 137, 164, 197, 237, 284, 341, 410, 492, 590, 708, 850, 1020

Table 5.1: Chord - finger distances for $m = 10$ ($id_{max} = 2^{10} - 1 = 1023$)

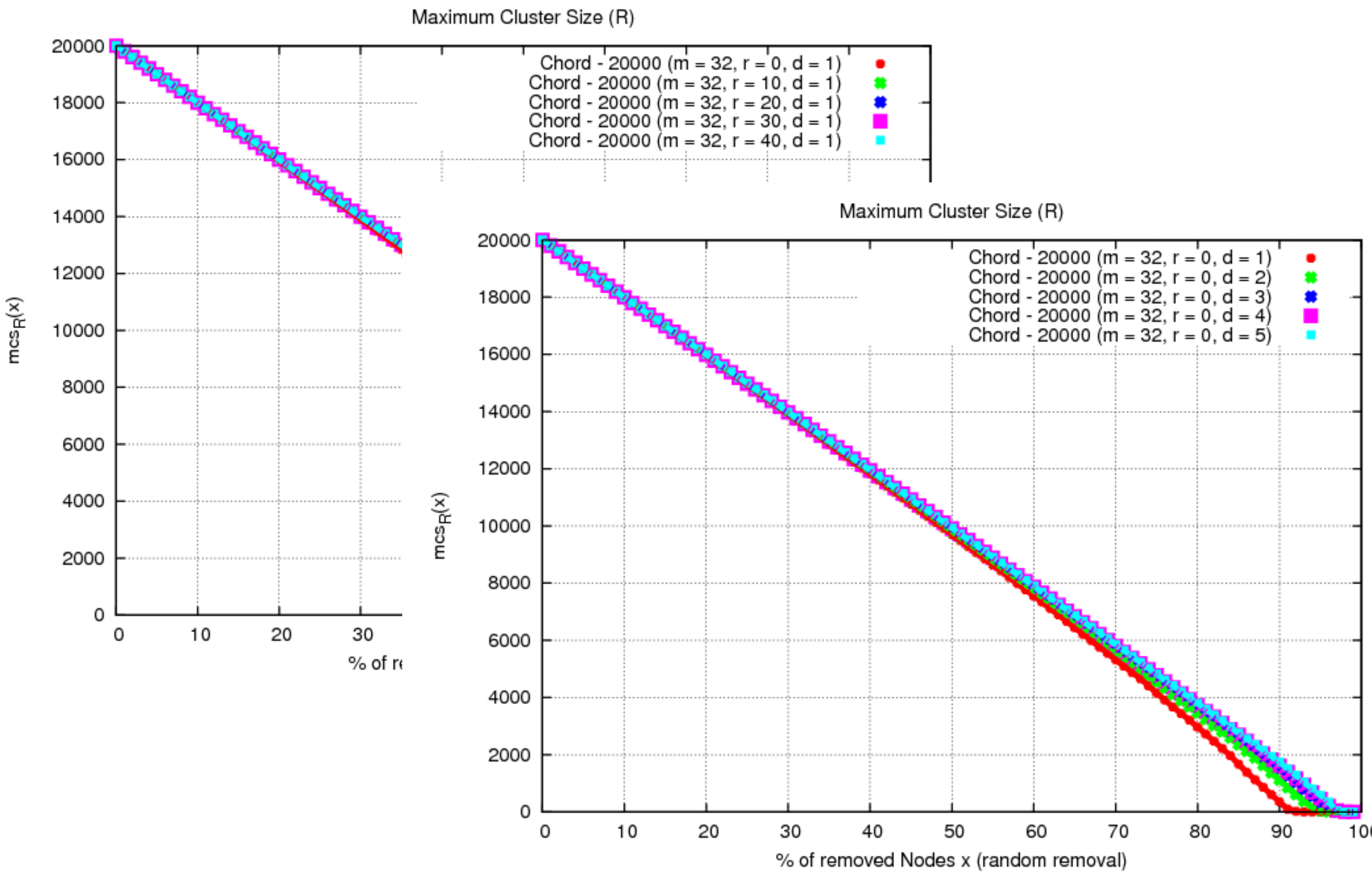
Analysing CHORD (3)



Analysing CHORD (4, Attack)



Analysing CHORD (5, Failure)



Analysing CHORD (6, Attack)