Web Information Retrieval

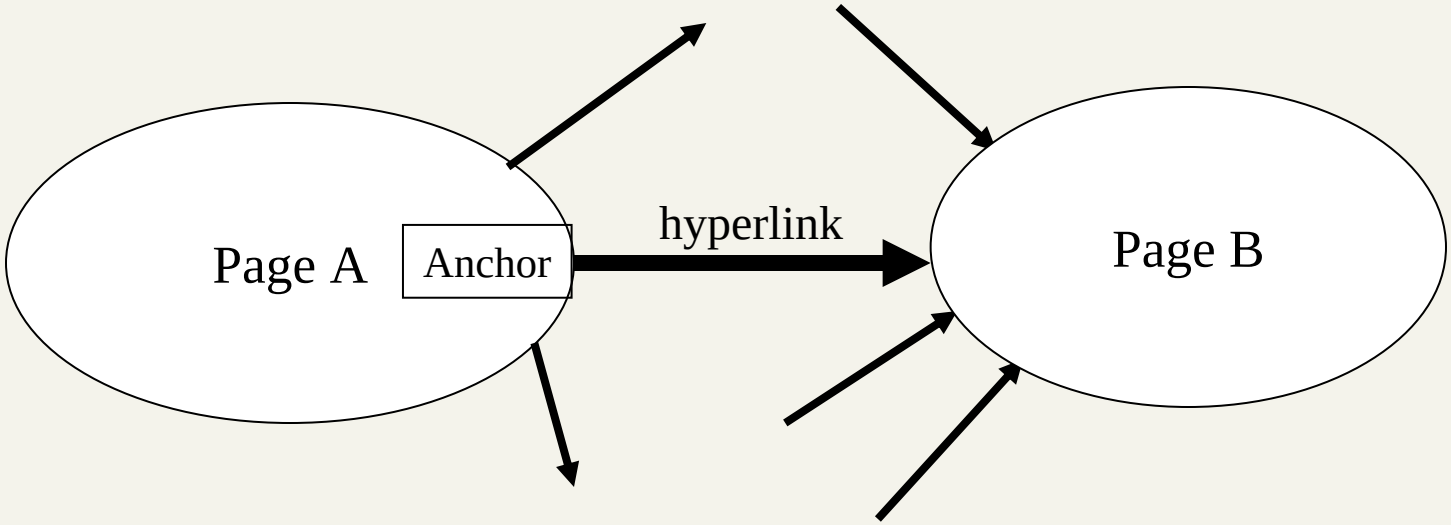
Lecture 11

Graph Structure of the Web for IR

Today’s lecture

- Anchor text
- Graph structure of the web

The Web as a Directed Graph

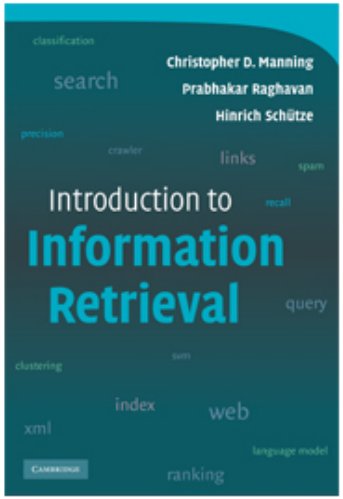


Assumption 1: A hyperlink between pages denotes author perceived relevance (quality signal)

Assumption 2: The anchor of the hyperlink describes the target page (textual context)

Assumption 1: reputed sites

Introduction to Information Retrieval



This is the companion website for the following book.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Informat*

You can order this book at CUP, at your local bookstore or on the internet. The best search

The book aims to provide a modern approach to information retrieval from a computer scie
University and at the University of Stuttgart.

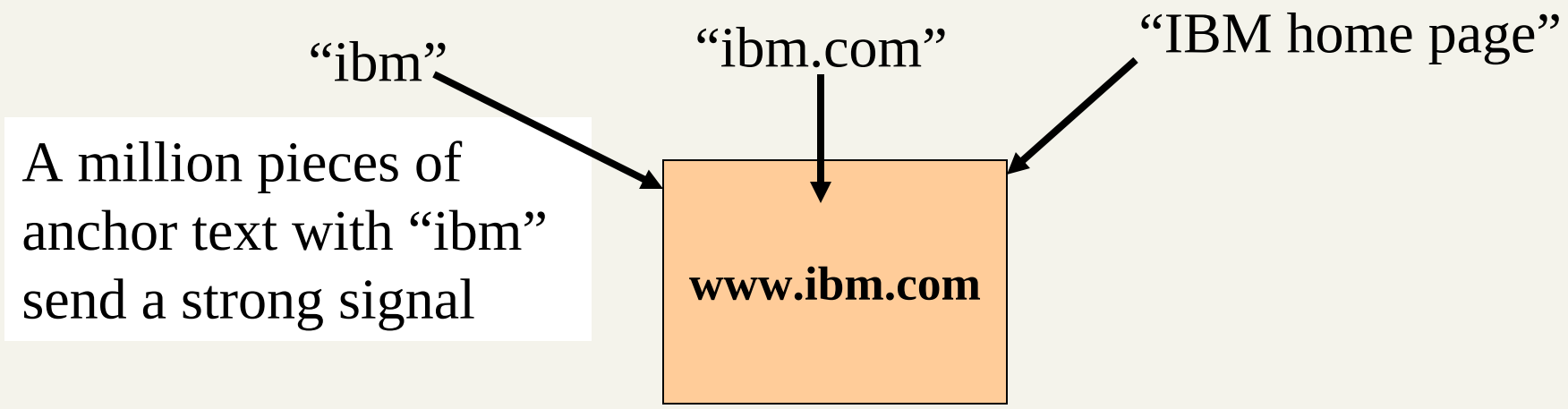
We'd be pleased to get feedback about how this book works out as a textbook, what is m
comments to: informationretrieval (at) yahoogroups (dot) com

Assumption 2: annotation of target

Anchor Text

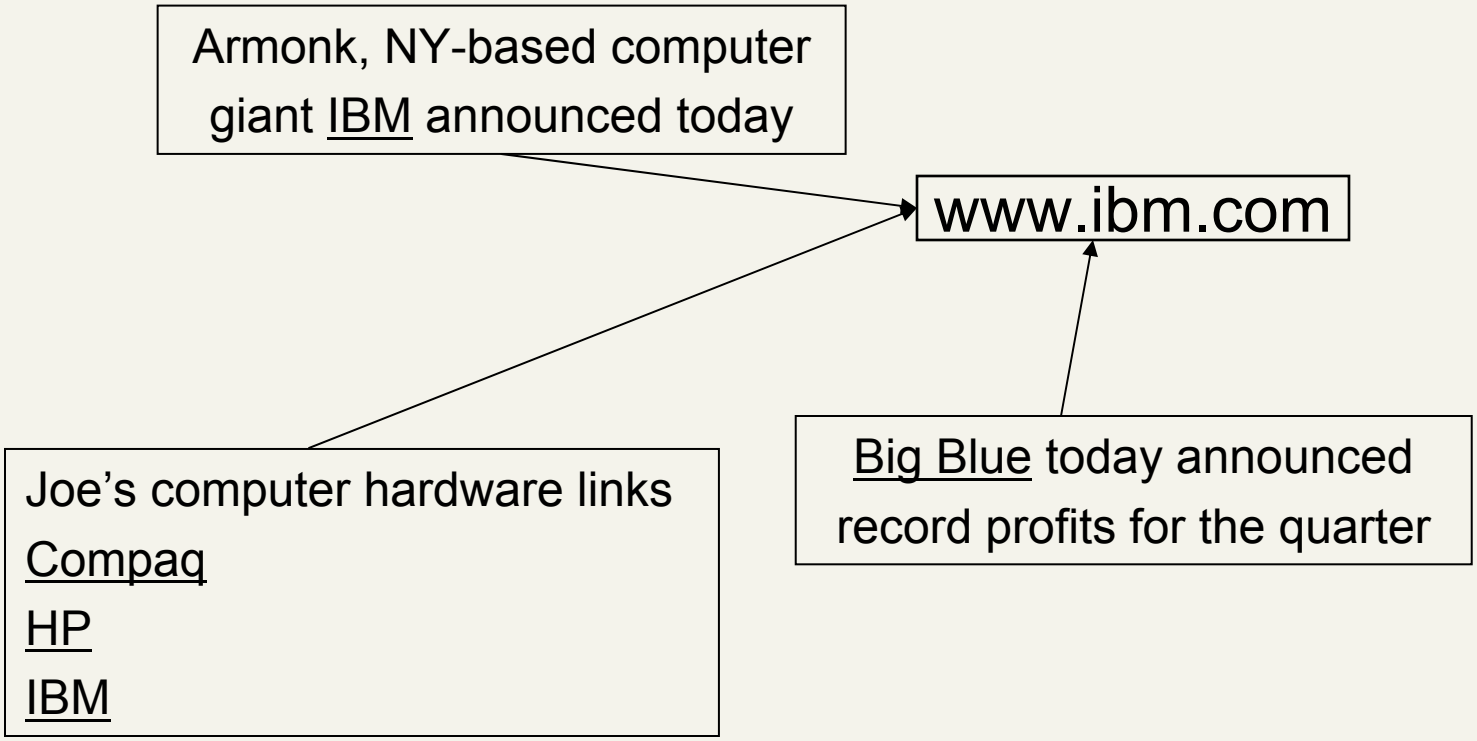
WWW Worm - McBryan [Mcbr94]

- For ***ibm*** how to distinguish between:
  - IBM's home page (mostly graphical)
  - IBM's copyright page (high term freq. for 'ibm')
  - Rival's spam page (arbitrarily high term freq.)



Indexing anchor text

- When indexing a document *D*, include anchor text from links pointing to *D*.



Indexing anchor text

- Can sometimes have unexpected side effects - *e.g.*, "***miserable failure***"
- Can score anchor text with weight depending on the authority of the anchor page's website
- E.g., if we were to assume that content from cnn.com or yahoo.com is authoritative, then trust the anchor text from them

Google Bombing

Anchor Text

- Other applications
  - Weighting/filtering links in the graph
    - HITS
  - Generating page descriptions from anchor text

The Web Graph

- Other applications

Web graph

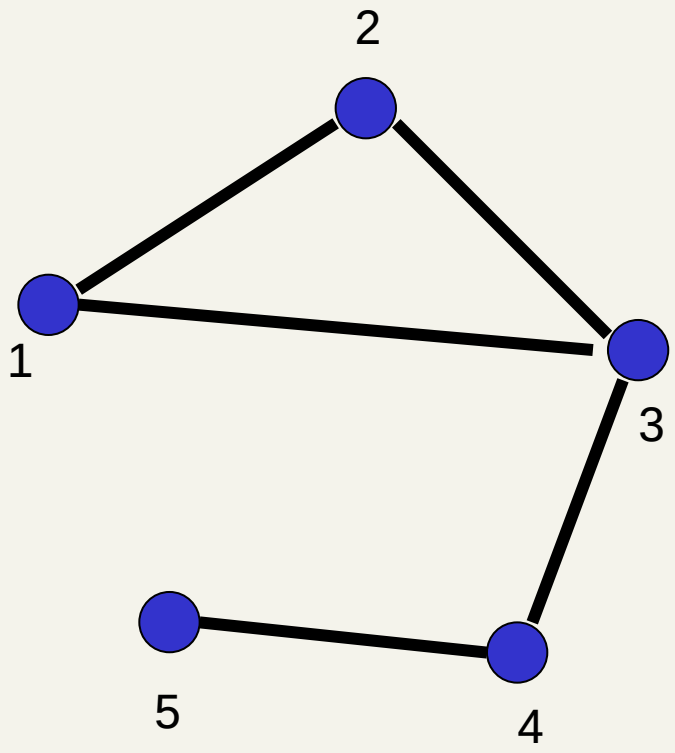
- Notation: G = (V, E) is given by
  - a set of vertices (nodes) denoted V
  - a set of edges (links) = pairs of nodes denoted E

- The page graph (directed)
  - V = static web pages (50 B)
  - E = static hyperlinks (350 B?)

pages indexed by Google on March 5

Graph Theory

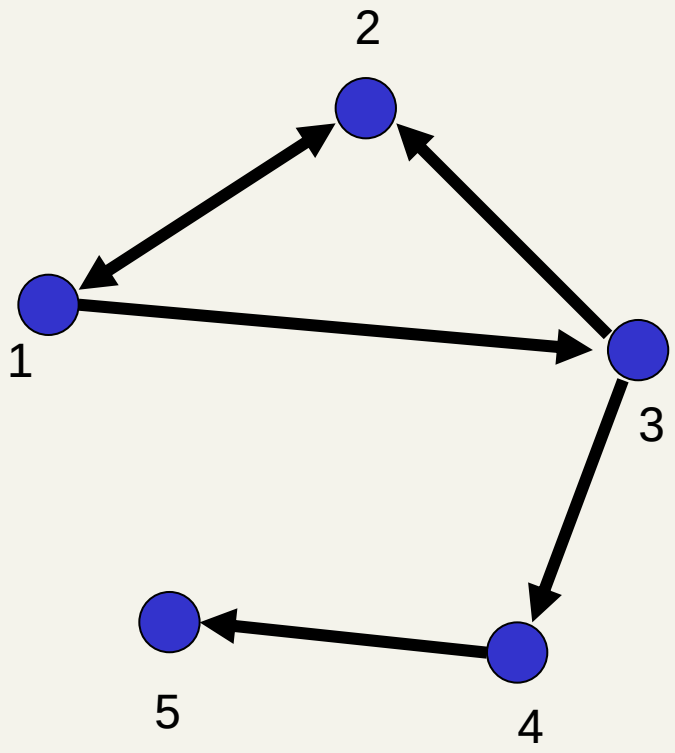
- Graph G=(V,E)
  - V = set of vertices
  - E = set of edges



undirected graph
E={(1,2),(1,3),(2,3),(3,4),(4,5)}

Graph Theory

- Graph G=(V,E)
  - V = set of vertices
  - E = set of edges

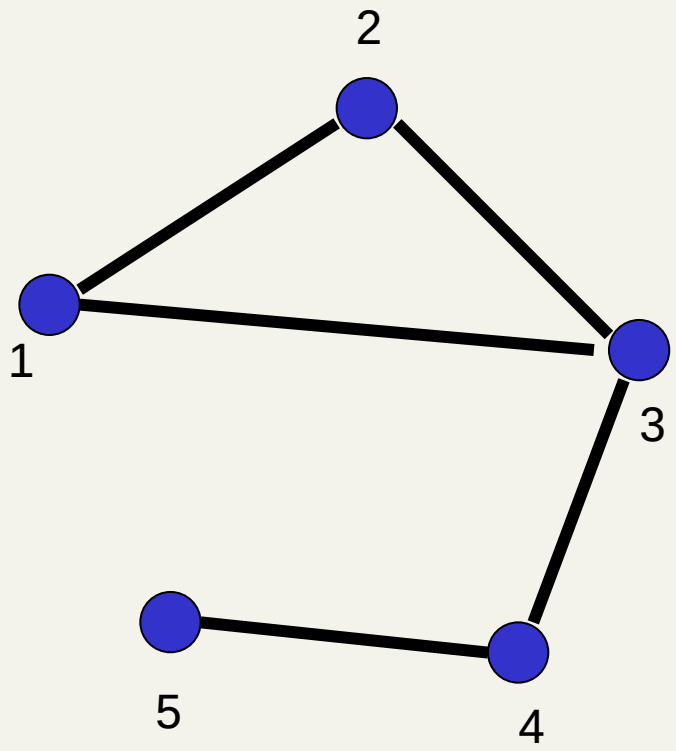


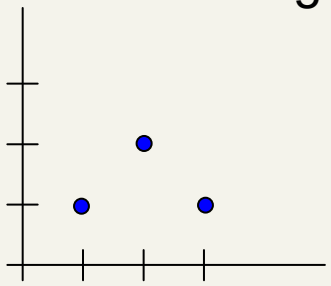
directed graph

E={‹1,2›, ‹2,1› ‹1,3›, ‹3,2›, ‹3,4›, ‹4,5›}

Undirected graph

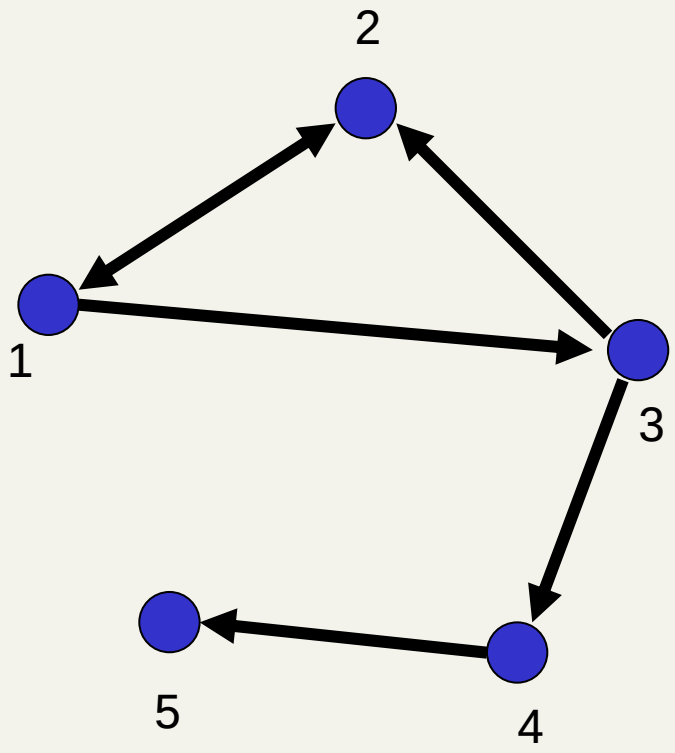
- degree d(i) of node i
  - number of edges incident on node i
- degree sequence
  - [d(1),d(2),d(3),d(4),d(5)]
  - [2,2,3,2,1]
- degree distribution
  - [(1,1),(2,3), (3,1)]





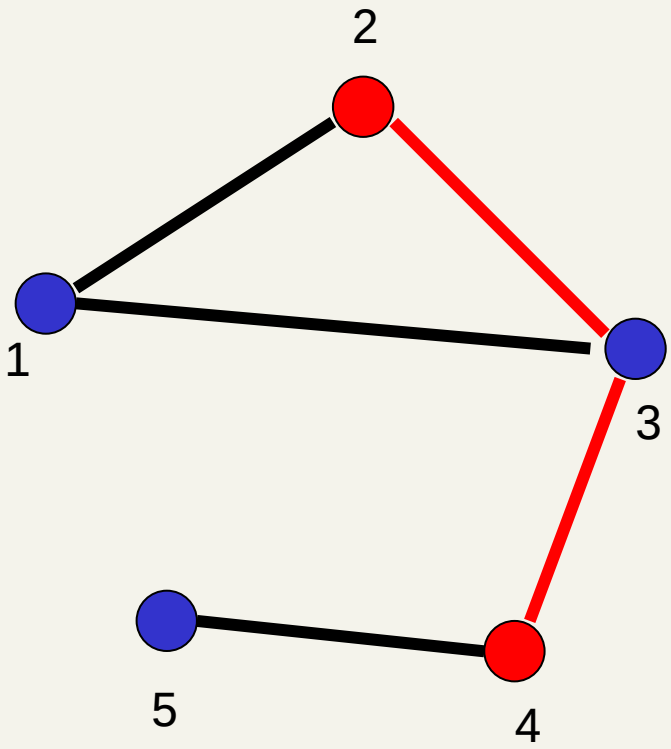
Directed Graph

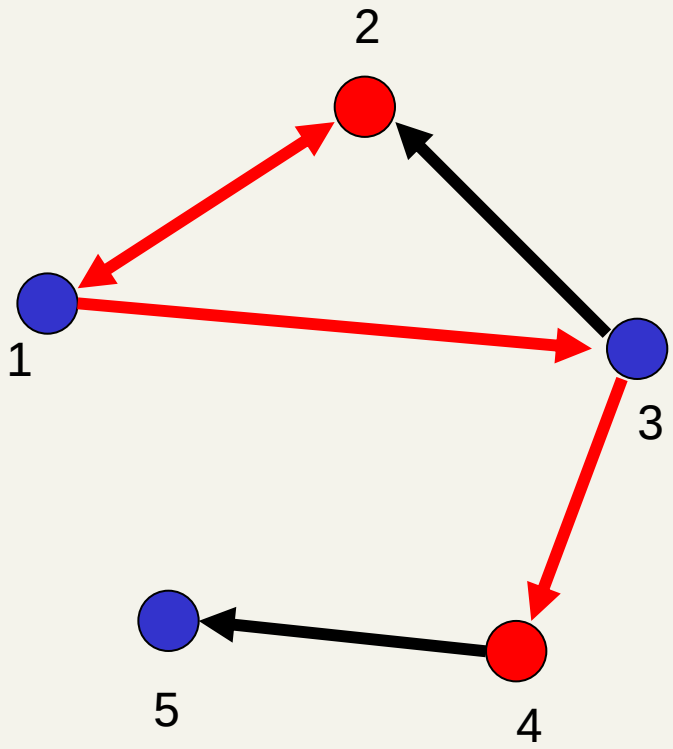
- in-degree $d_{in}(i)$ of node i
  - number of edges pointing to node i
- out-degree $d_{out}(i)$ of node i
  - number of edges leaving node i
- in-degree sequence
  - [1,2,1,1,1]
- out-degree sequence
  - [2,1,2,1,0]



Paths

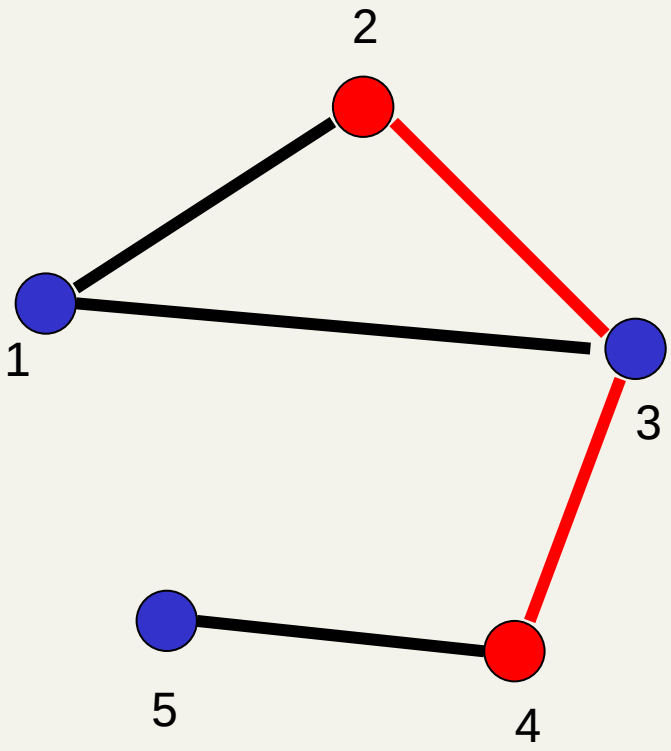
- Path from node i to node j: a sequence of edges (directed or undirected from node i to node j)
  - path length: number of edges on the path
  - nodes i and j are connected

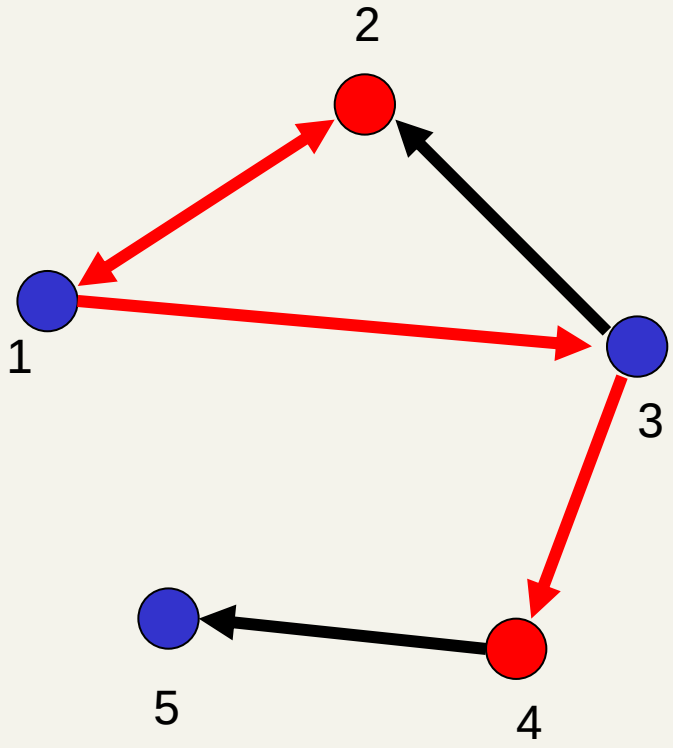




Shortest Paths

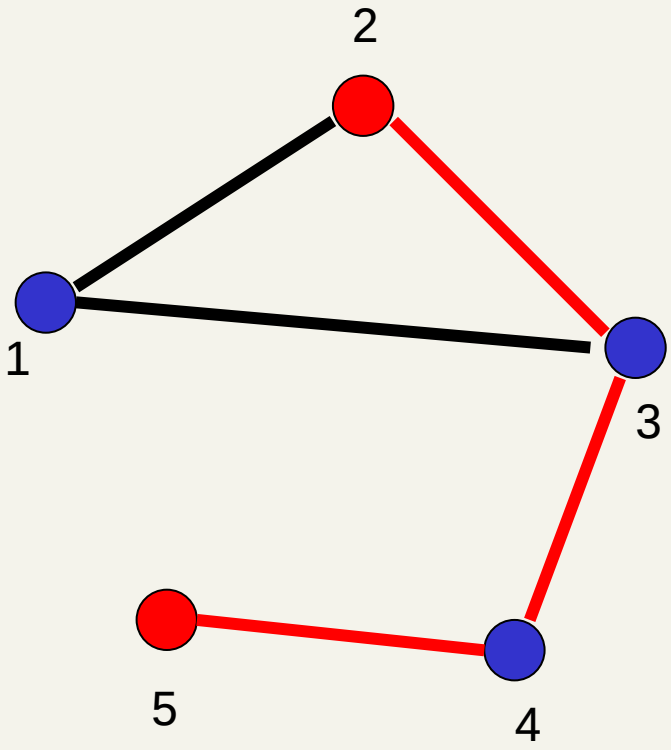
- Shortest Path from node i to node j
  - also known as BFS path, or geodesic path

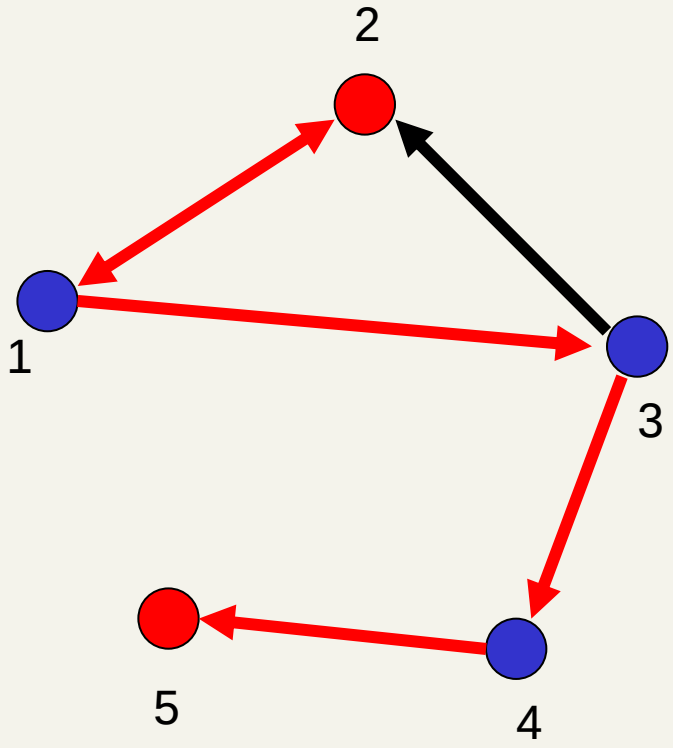




Diameter

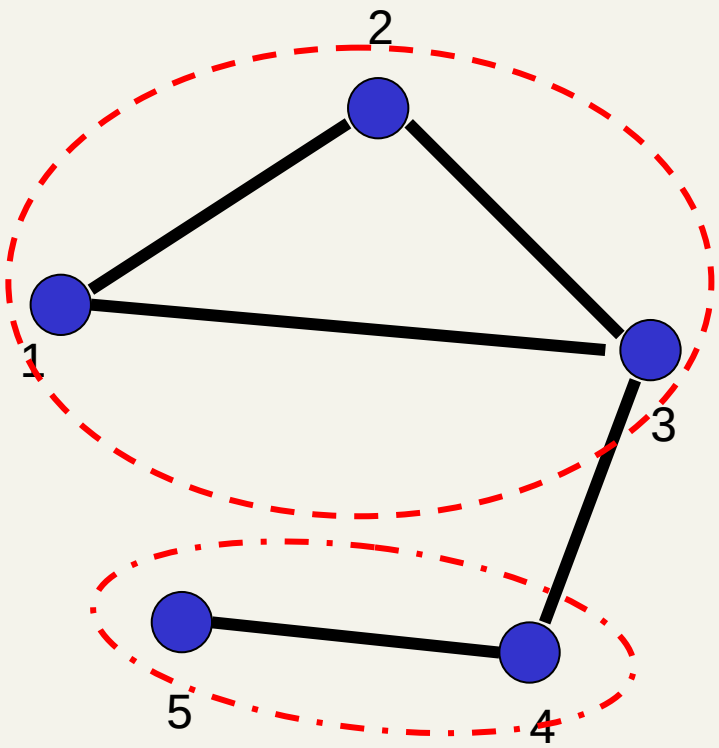
- The longest shortest path in the graph





Undirected graph

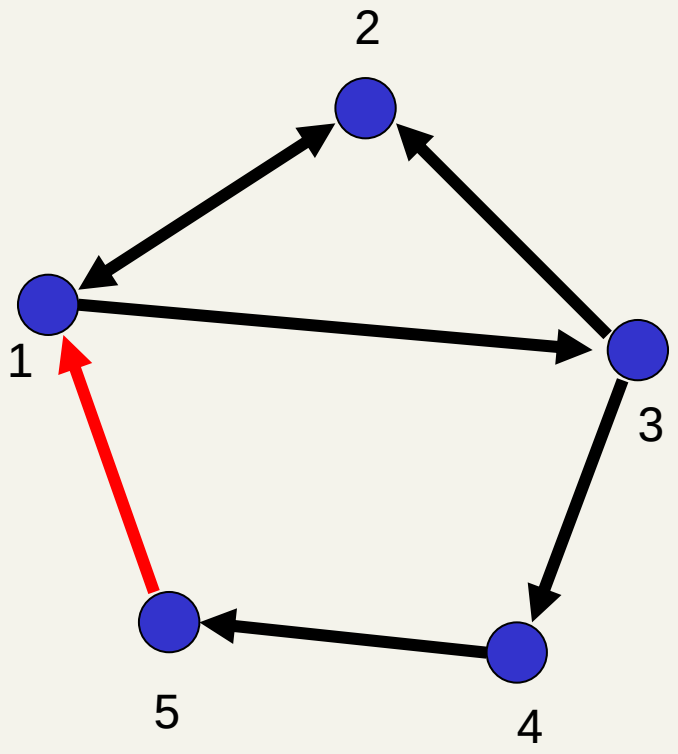
- Connected graph: a graph where there every pair of nodes is connected
- Disconnected graph: a graph that is not connected
- Connected Components: subsets of vertices that are connected



Directed Graph

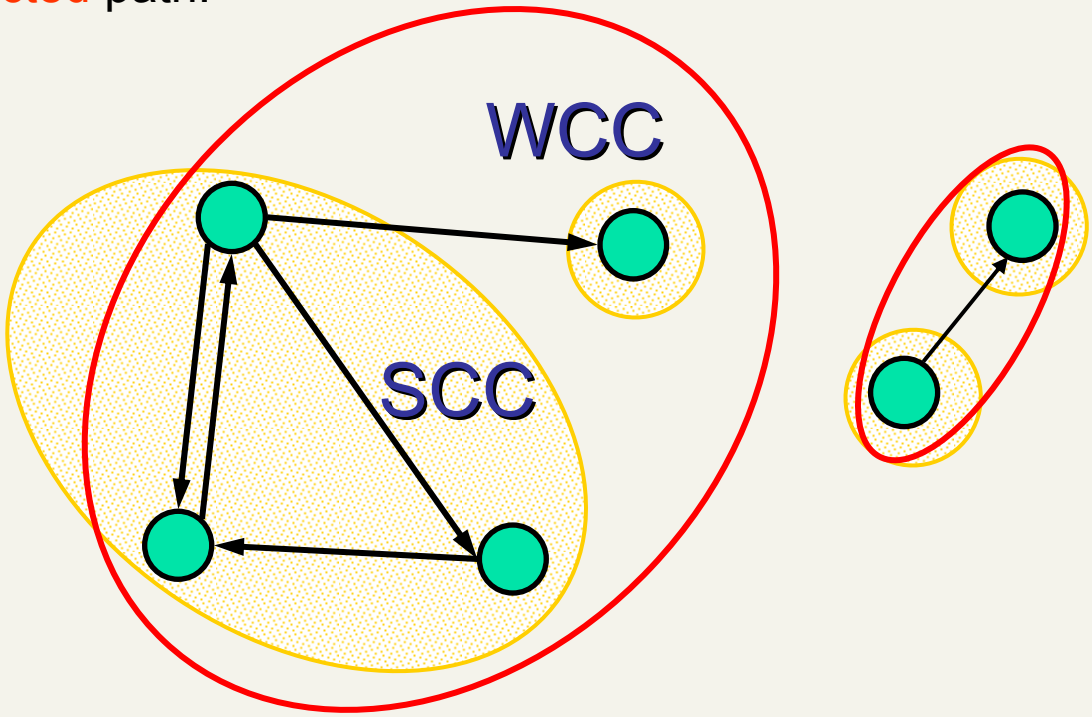
- Strongly connected graph: there exists a path from every i to every j

- Weakly connected graph: If edges are made to be undirected the graph is connected



Connected components – definitions

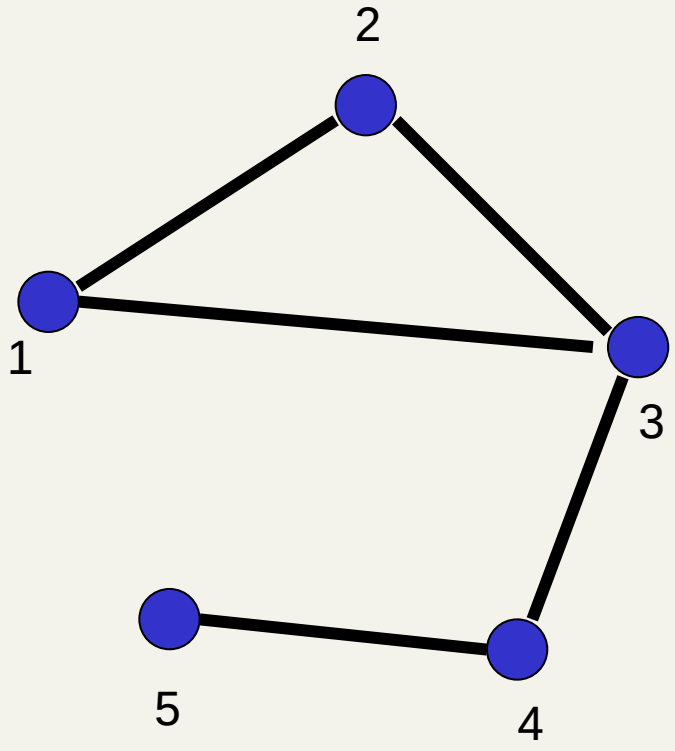
- Weakly connected components (WCC)
  - Set of nodes such that from any node can go to any node via an undirected path
- Strongly connected components (SCC)
  - Set of nodes such that from any node can go to any node via a directed path.



Adjacency matrix

- Adjacency matrix
  - symmetric matrix for undirected graphs

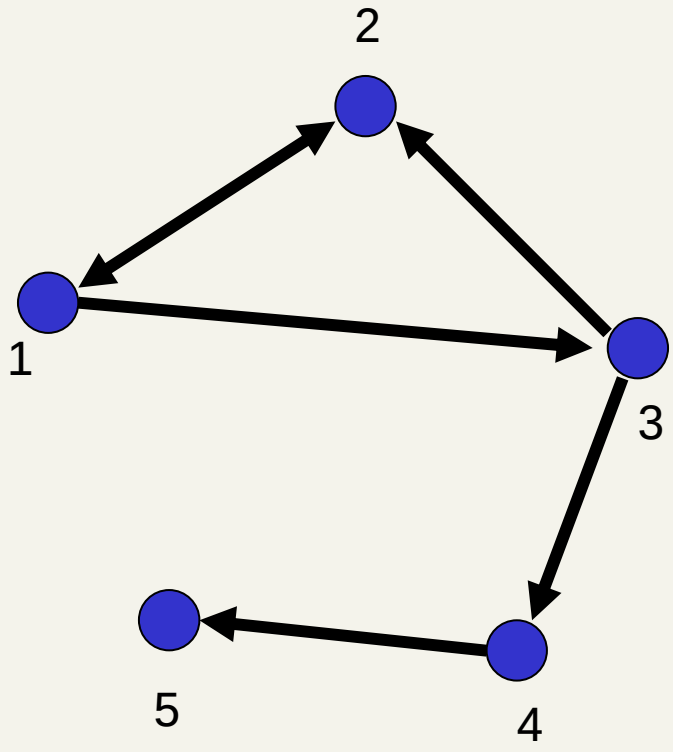
$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$



Adjacency matrix

- Adjacency matrix
  - unsymmetric matrix for undirected graphs

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



Why is it interesting to study the Web Graph?

- It is the largest artifact ever conceived by the human
- Exploit its structure of the Web for
  - Crawl strategies
  - Search
  - Spam detection
  - Discovering communities on the web
  - Classification/organization
- Predict the evolution of the Web
  - Mathematical models
  - Sociological understanding

Many other web/internet related graphs

- Physical network graph
  - V = Routers
  - E = communication links
- The host graph (directed)
  - V = hosts
  - E = There is an edge from a page on host A to a page on host B
- The “cosine” graph (undirected, weighted)
  - V = static web pages
  - E = cosine distance among term vectors associated with pages
- Co-citation graph (undirected, weighted)
  - V = static web pages
  - E = (x,y) number of pages that refer to both x and y
- Communication graphs (which hosts talks to which hosts at a given time)
- Routing graph (how packets move)
- Social networks
- Biology
- …

Observing Web Graph

- It is a huge ever expanding graph
- We do not know which percentage of it we know
- The only way to discover the graph structure of the web as hypertext is via large scale crawls

- Warning: the picture might be distorted by
  - Size limitation of the crawl
  - Crawling rules
  - Perturbations of the "natural" process of birth and death of nodes and links

Naïve solution

- Keep crawling, when you stop seeing more pages, stop
- Extremely simple but wrong solution: crawling is complicated because the web is complicated
  - spamming
  - duplicates
  - mirrors
- First example of a complication: Soft 404
  - When a page does not exists, the server is supposed to return an error code = “404”
  - Many servers do not return an error code, but keep the visitor on site, or simply send him to the home page

The Static Public Web

- Static
  - not the result of a cgi-bin scripts
  - no “?” in the URL
  - doesn’t change very often
- Public
  - no password required
  - no robots.txt exclusion
  - no “noindex” meta tag

Graph properties of the Web

- **Global structure:** how does the web look from far away ?
- **Connectivity:** how many connections?
- **Connected components:** how large?
- **Reachability:** can one go from here to there ? How many hops?
- **Dense Subgraphs:** are there good clusters?

Power laws

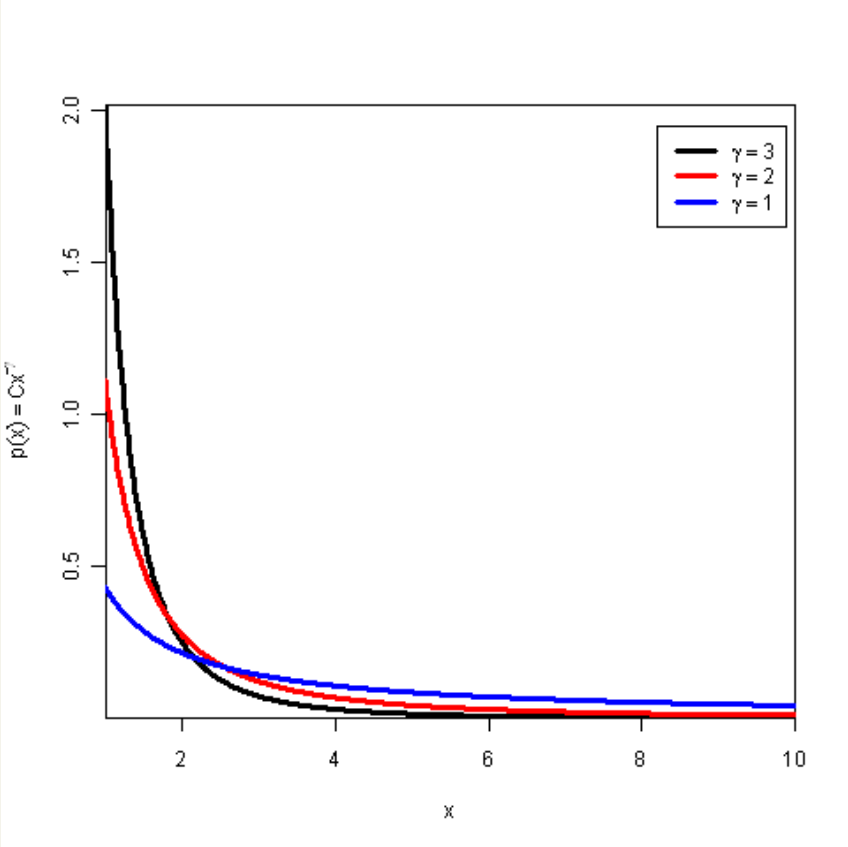
- Inverse polynomial tail (for large k)
  - $\Pr(X = k) \sim 1/k^{\alpha}$
  - Graph of such a function is a line on a log-log scale

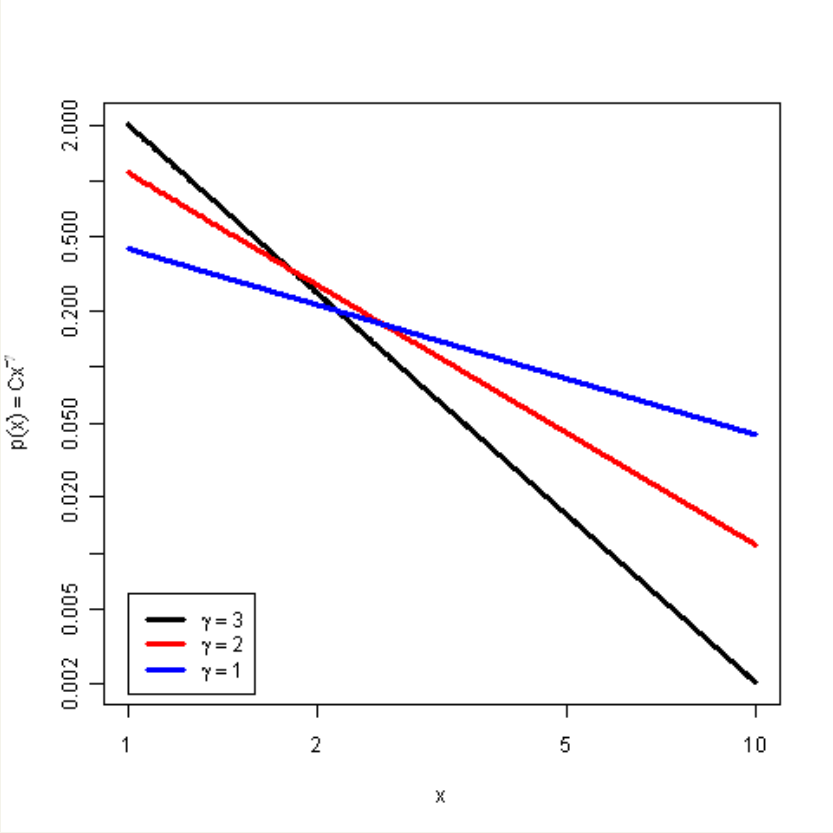
    $$\log(\Pr(X=k)) = -\alpha * \log(k)$$
  - Very large values possible & probable

- Exponential tail (for large k)
  - $\Pr(X = k) \sim e^{-k}$
  - Graph of such function is a line on a log scale

    $$\log(\Pr(X=k)) = -k$$
  - No very large values

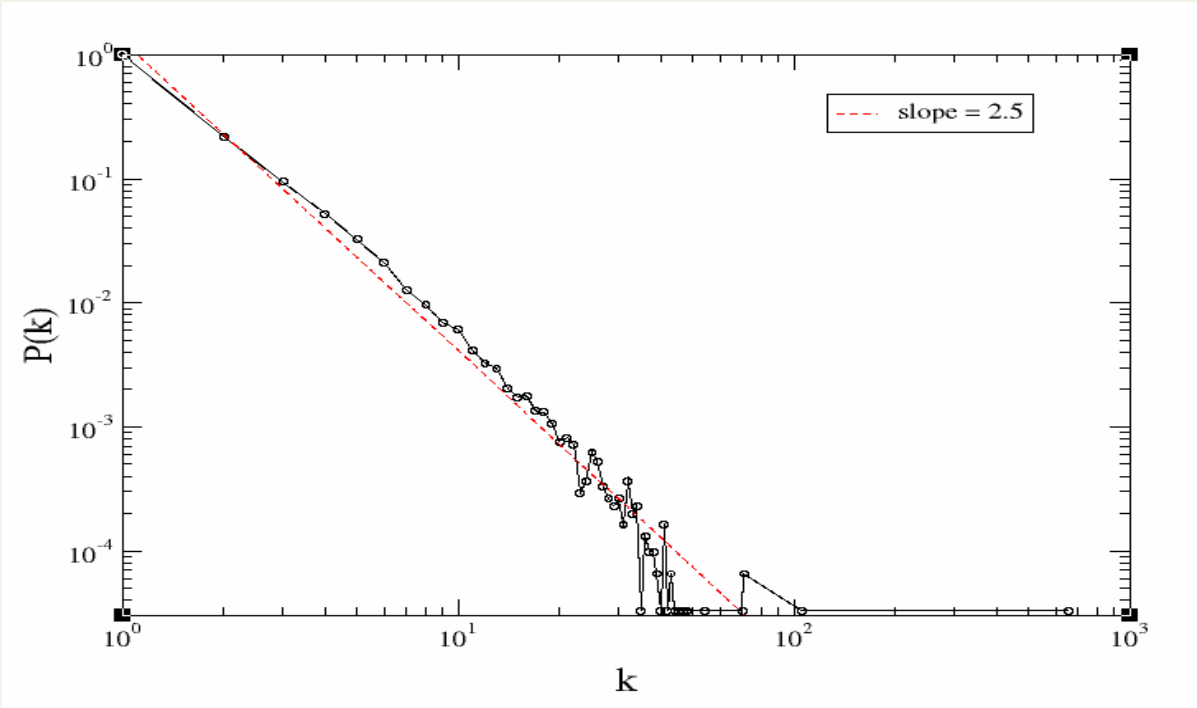
Power laws


p(x) = Cx^-γ
2.0
1.5
1.0
0.5
γ = 3
γ = 2
γ = 1
2
4
6
8
10
x


p(x) = Cx^-γ
2.000
0.500
0.200
0.050
0.020
0.005
0.002
γ = 3
γ = 2
γ = 1
1
2
5
10
x

Power laws

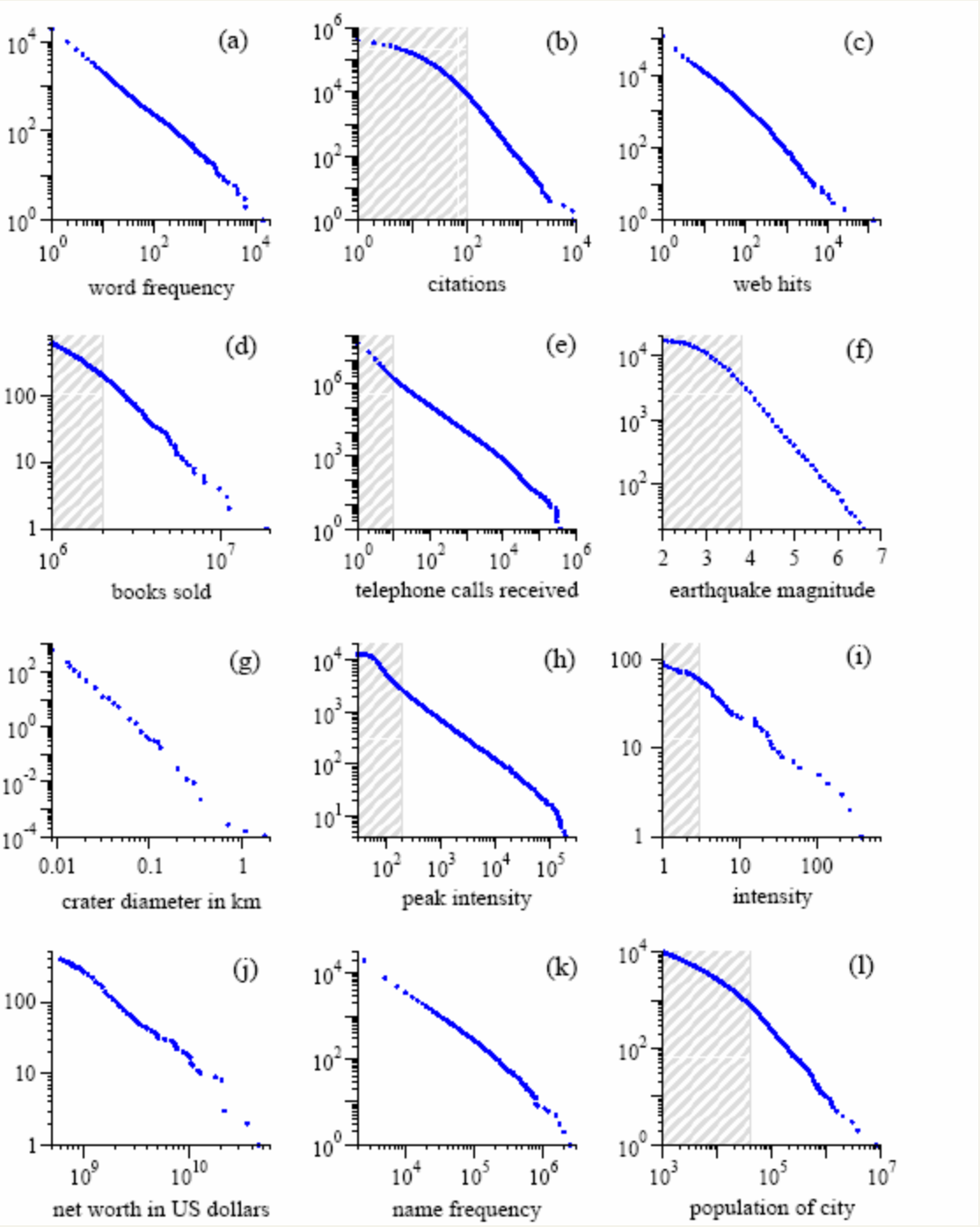
- Examples
  - Inverse poly tail: the distribution of wealth, popularity of books, etc.
  - Inverse exp tail: the distribution of age, etc
  - Internet network topology [Faloutsos & al. 99]



Power laws are ubiquitous

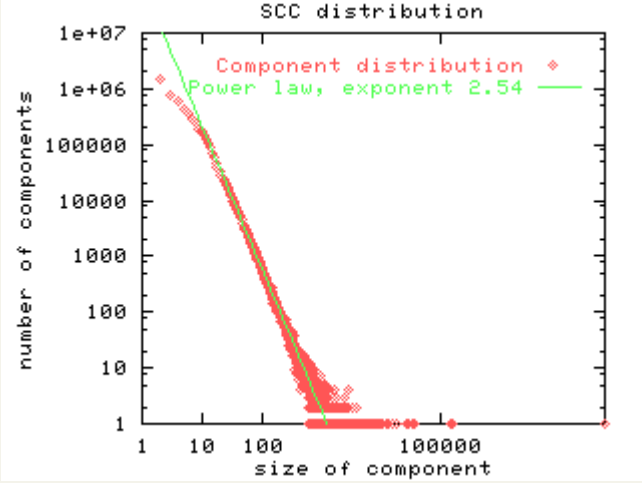
	quantity	minimum $x_{\min}$	exponent α
(a)	frequency of use of words	1	2.20(1)
(b)	number of citations to papers	100	3.04(2)
(c)	number of hits on web sites	1	2.40(1)
(d)	copies of books sold in the US	2 000 000	3.51(16)
(e)	telephone calls received	10	2.22(1)
(f)	magnitude of earthquakes	3.8	3.04(4)
(g)	diameter of moon craters	0.01	3.14(5)
(h)	intensity of solar flares	200	1.83(2)
(i)	intensity of wars	3	1.80(9)
(j)	net worth of Americans	\$600m	2.09(4)
(k)	frequency of family names	10 000	1.94(1)
(l)	population of US cities	40 000	2.30(5)

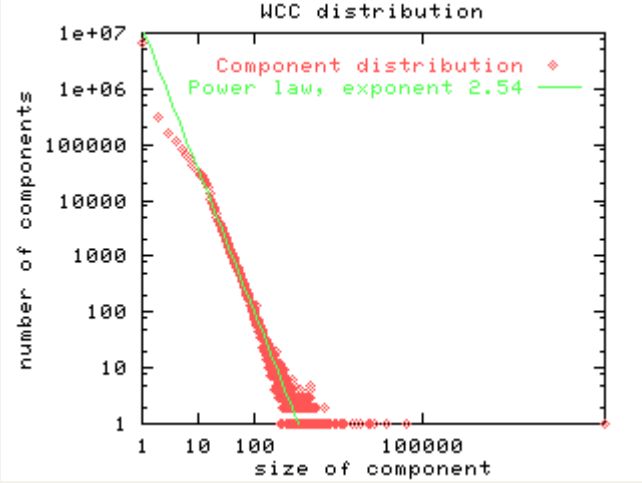
TABLE I Parameters for the distributions shown in Fig. 4. The labels on the left refer to the panels in the figure. Exponent values were calculated using the maximum likelihood method of Eq. (5) and Appendix B, except for the moon craters (g), for which only cumulative data were available. For this case the exponent quoted is from a simple least-squares fit and should be treated with caution. Numbers in parentheses give the standard error on the trailing figures.



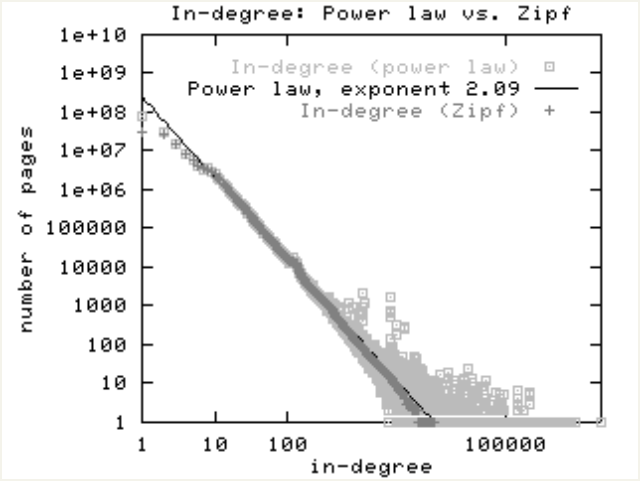
SCC and WCC distribution

- The SCC and WCC sizes follows a power law distribution
  - the second largest SCC is significantly smaller

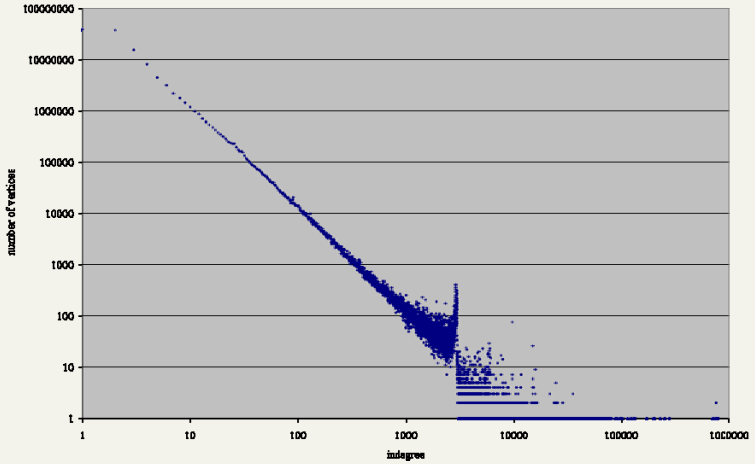




The In-degree distribution



Altavista crawl, 1999



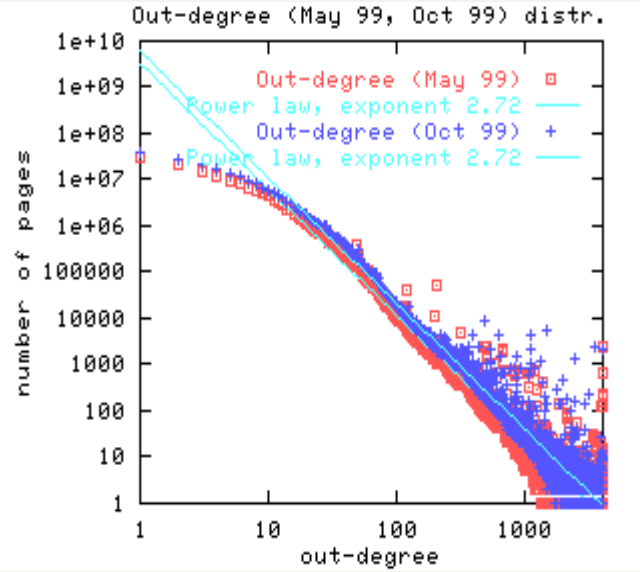
WebBase Crawl 2001

Indegree follows power law distribution

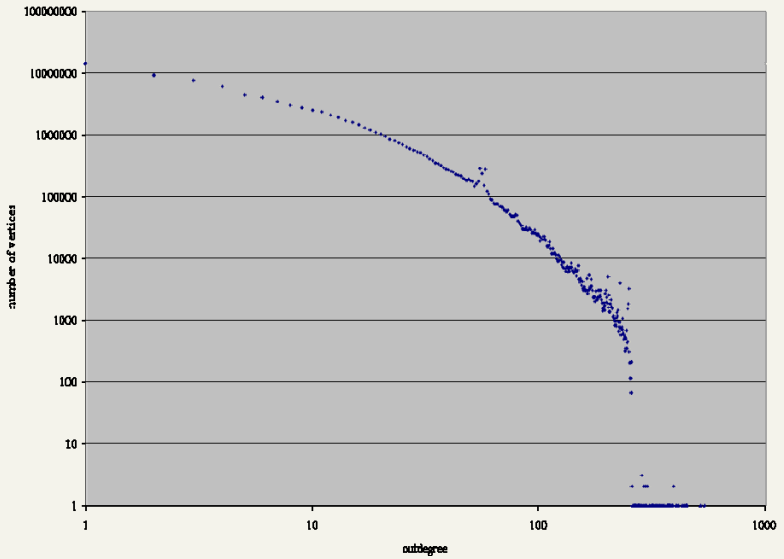
$$\Pr[\text{in-degree}(u) = k] \propto \frac{1}{k^{\alpha}}$$

$$\alpha = 2.1$$

Graph structure of the Web



Altavista, 1999



WebBase 2001

Out-degree follow power law distribution? Pages with a large number of outlinks are less frequent.

What is good about 2.1

- Expected number of nodes with degree k

$$\sim n / k^{2.1}$$

- Expected contribution to total edges

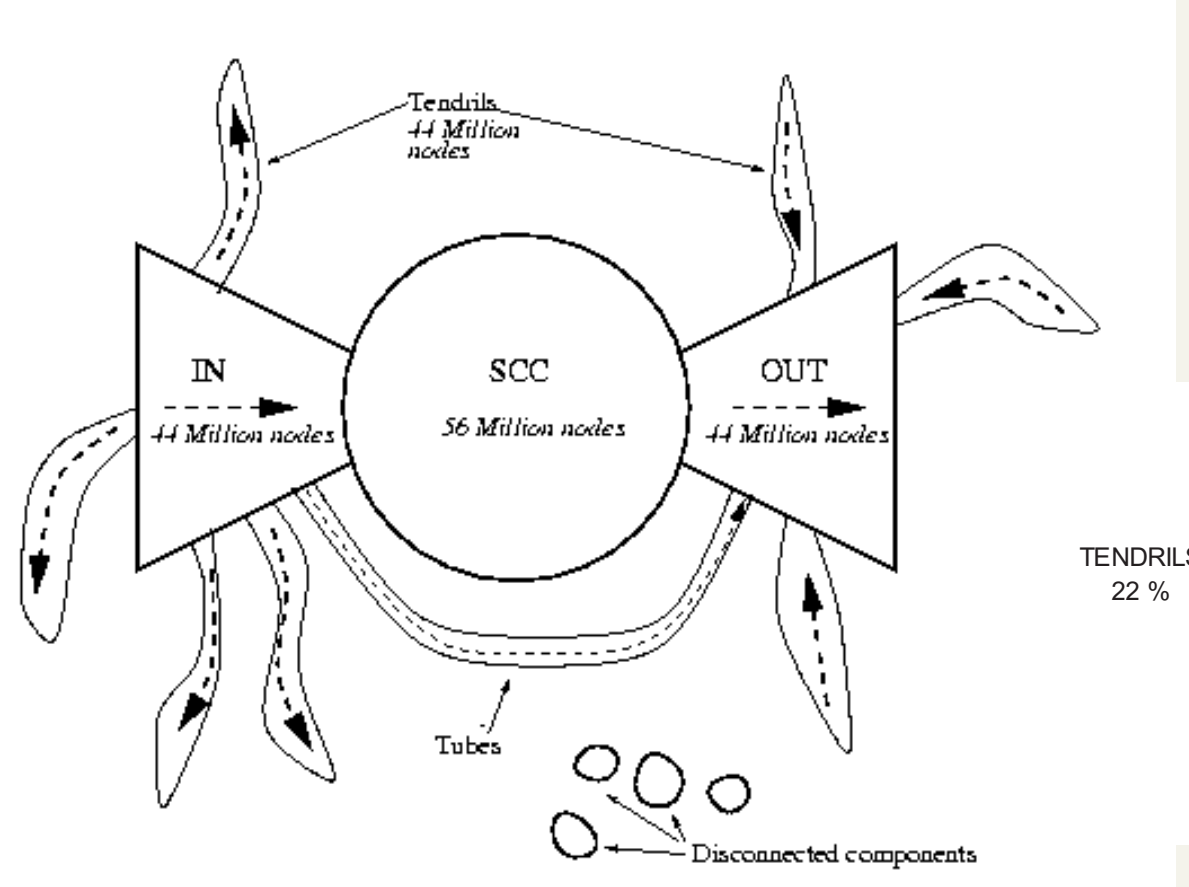
$$\sim n / k^{2.1} * k = \sim n / k^{1.1}$$

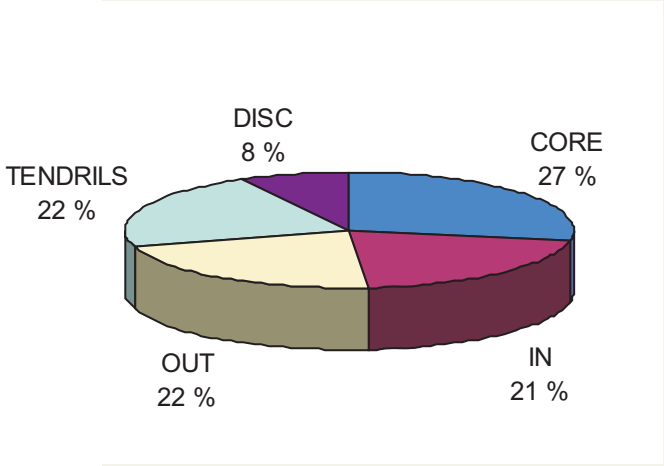
- Summing over all k, the expected number of edges is

$$\sim n$$

- This is good news: our computational power is such that we can deal with work linear in the size of the web, but not much about that.

The bow-tie structure of the Web



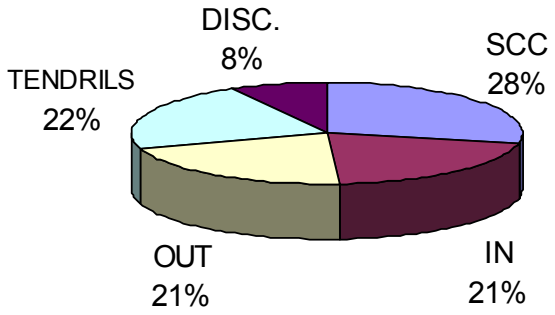


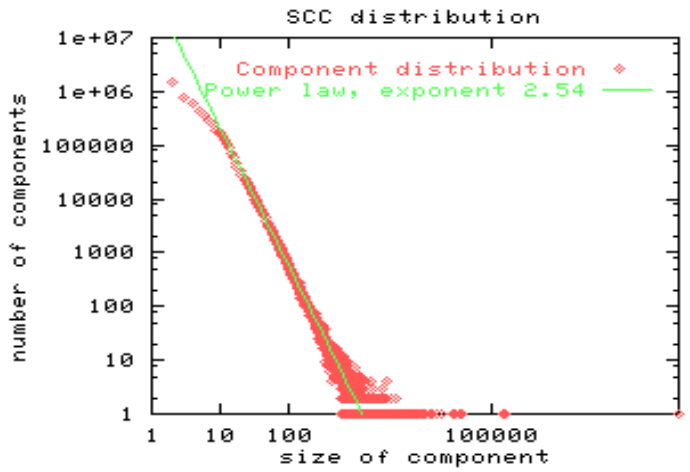
Papillon

- inglese: bow-tie
- italiano: papillon, farfallino,
- francese: noeud de papillon
- spagnolo: corbata, pajarita, …
- greco: παπιγιόν
- tedesco: fliege
- polacco: mucha

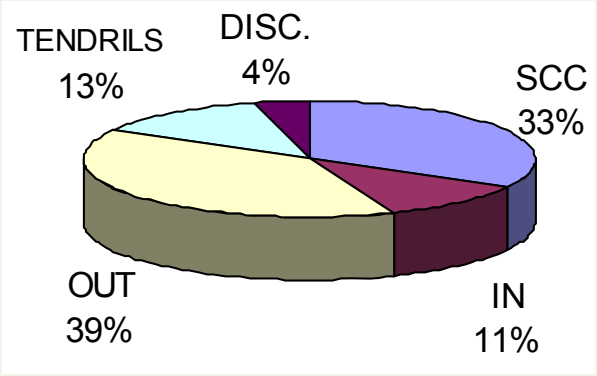
Experiments (1)

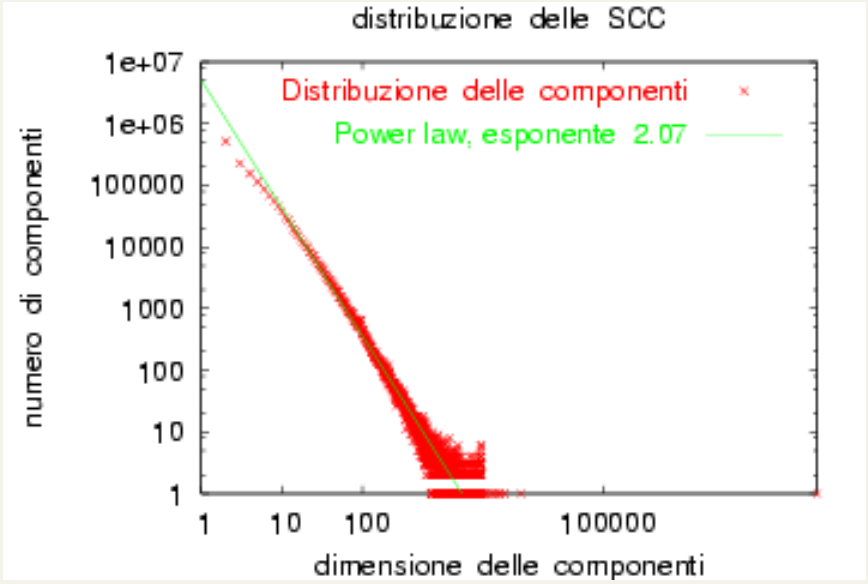
Altavista ‘99



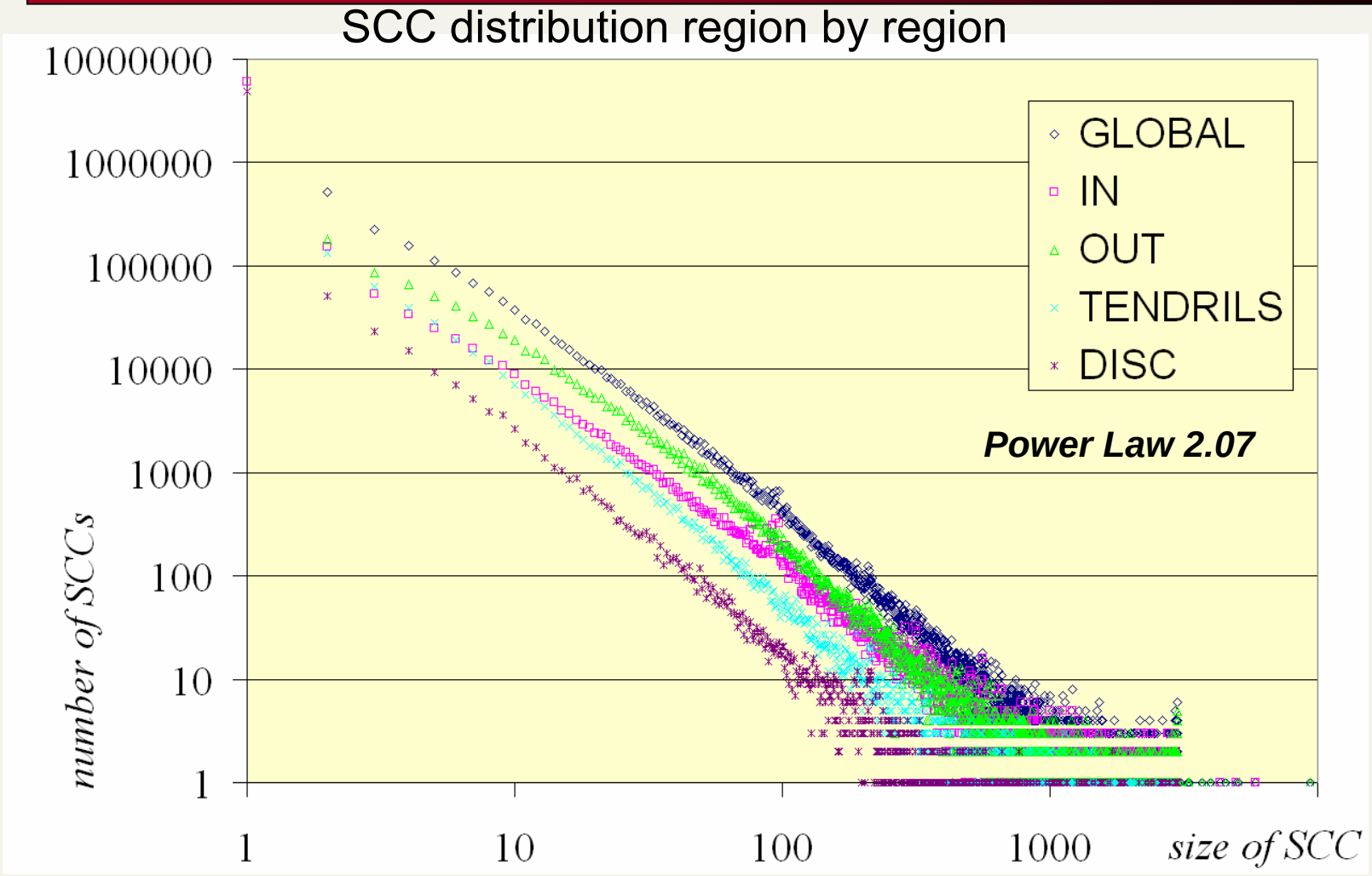


WebBase ‘01



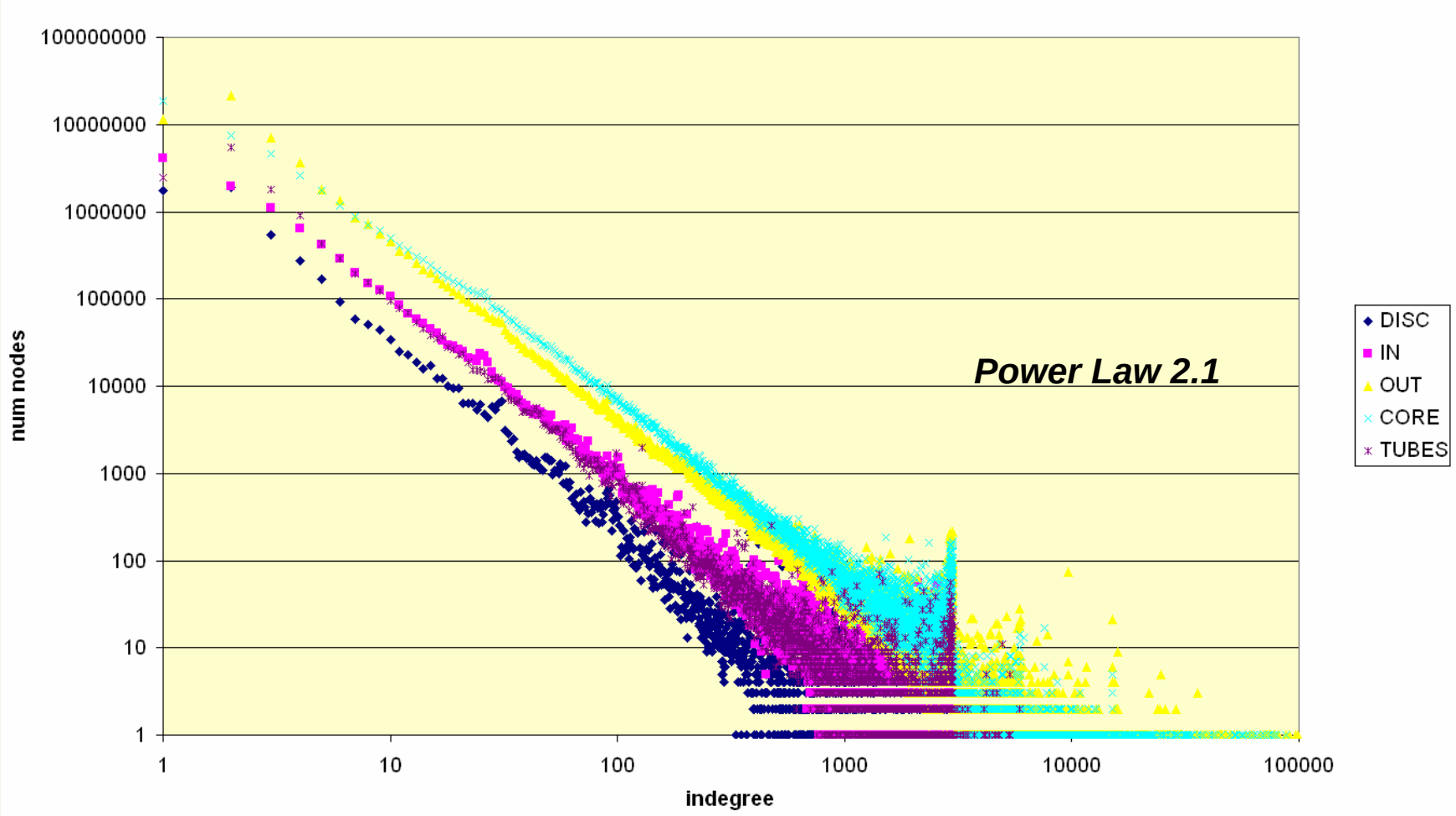


Experiments (2)



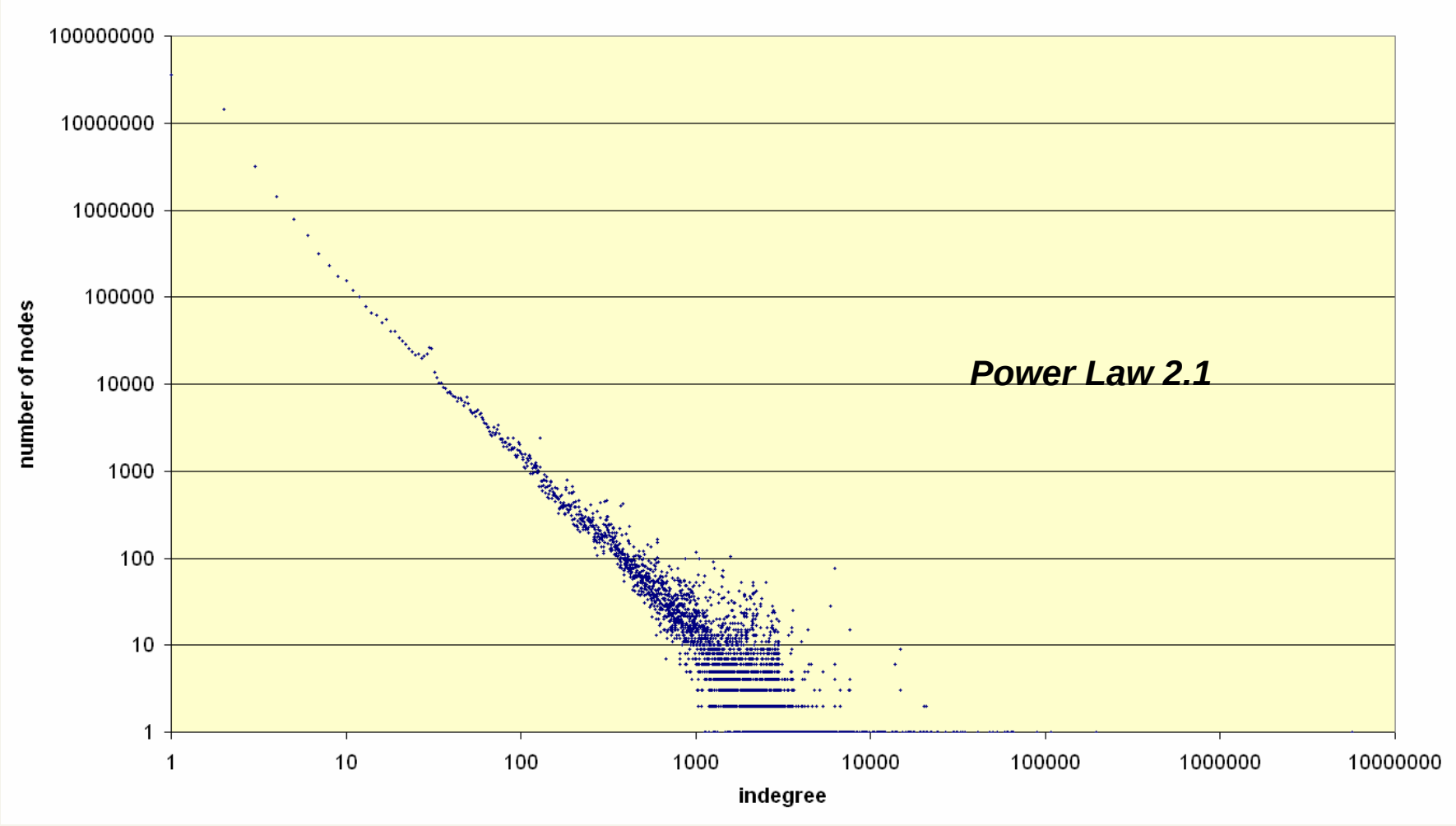
Experiments (3)

Indegree distribution region by region



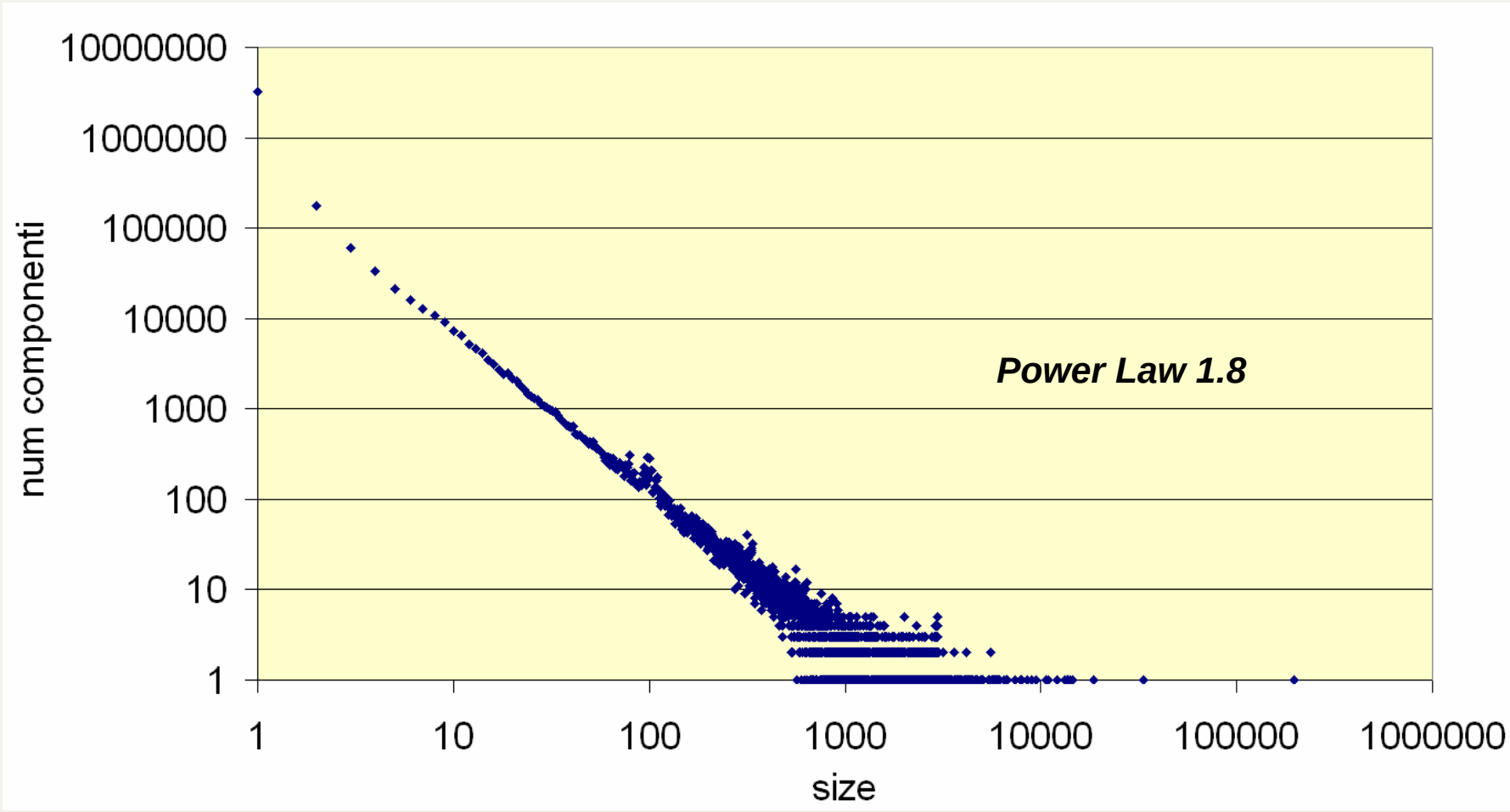
Experiments (4)

Indegree distribution in the SCC graph



Experiments (5)

WCC distribution in IN



What did we learn from all these power laws?

- The second largest SCC is of size less 10000 nodes!
- IN and Out have millions of access points to the CORE and thousands of relatively large Weakly Connected Components
- This may help in designing better crawling strategies at least for IN and OUT, i.e. the load can be splitted between the robots without much overlapping.
- While power law with exponent 2.1 is a universal feature of the Web, there is no fractal structure: IN and OUT do not show any Bow Tie Phenomena

Algorithmic issues

- Apply standard linear-time algorithms for WCC and SCC
  - Hard to do if you can't store the entire graph in memory!!
    - WCC is easy if you can store V (semi-external algorithm)
    - No one knows how to do DFS in semi-external memory, so SCC is hard ??
    - Might be able to do approx SCC, based on low diameter.
- Random sampling for connectivity information
  - Find all nodes reachable from one given node ("Breadth-First Search")
  - BFS is also hard. Simpler on low diameter

Find the CORE

- Iterate the following process:
  - Pick a random vertex **v**
  - Compute all node reached from **v**: O(**v**)
  - Compute all nodes that reach **v**: I(**v**)
  - Compute SCC(**v**):= I(**v**) ∩ O(**v**)
  - Check whether it is the largest SCC

If the CORE is about ¼ of the vertices, after 20 iterations, probability to not find the core < 1%.

Resources

- IIR Chapters 21 – 21.1