How does the Web look like?

CS 322: (Social and Information) Network Analysis

Jure Leskovec
Stanford University

Information networks

- Nodes are pieces of information
- Links join the pieces that are related to each other in some fashion
- Examples:
  - World Wide Web
  - Citation Networks
  - References in an encyclopedia
  - Internet
  - Wireless communication

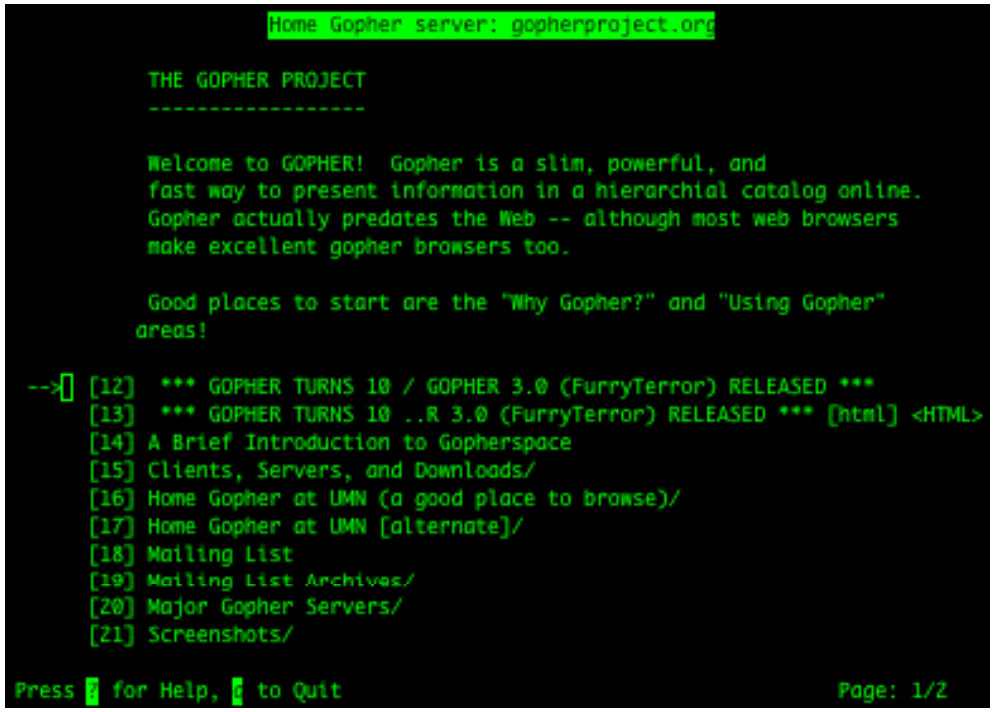
Before the Web was born

- Memex [*As we may think* by V. Bush, 1945]
  - Observed that information in books is highly linear
  - Our associative memory represents semantic network
  - Memex is a system that mimics such semantic networks with knowledge and links between the concepts

Before the Web was born

Gopher:

- Similar to Web:
  - links
- But:
  - No graphics
  - Imposes much stronger hierarchy – web of menu items – like a file system
  - Terminal based
  - “go for” (url)



1990: The Web is born

- Created by Tim Barners-Lee in 1990 in CERN:
  - a large hypertext database with typed links

Christmas 1990: First Web page

World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.

Help
on the browser you are using

Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)

Technical
Details of protocols, formats, program internals etc

Bibliography
Paper documentation on W3 and references.

People
A list of some people involved in the project.

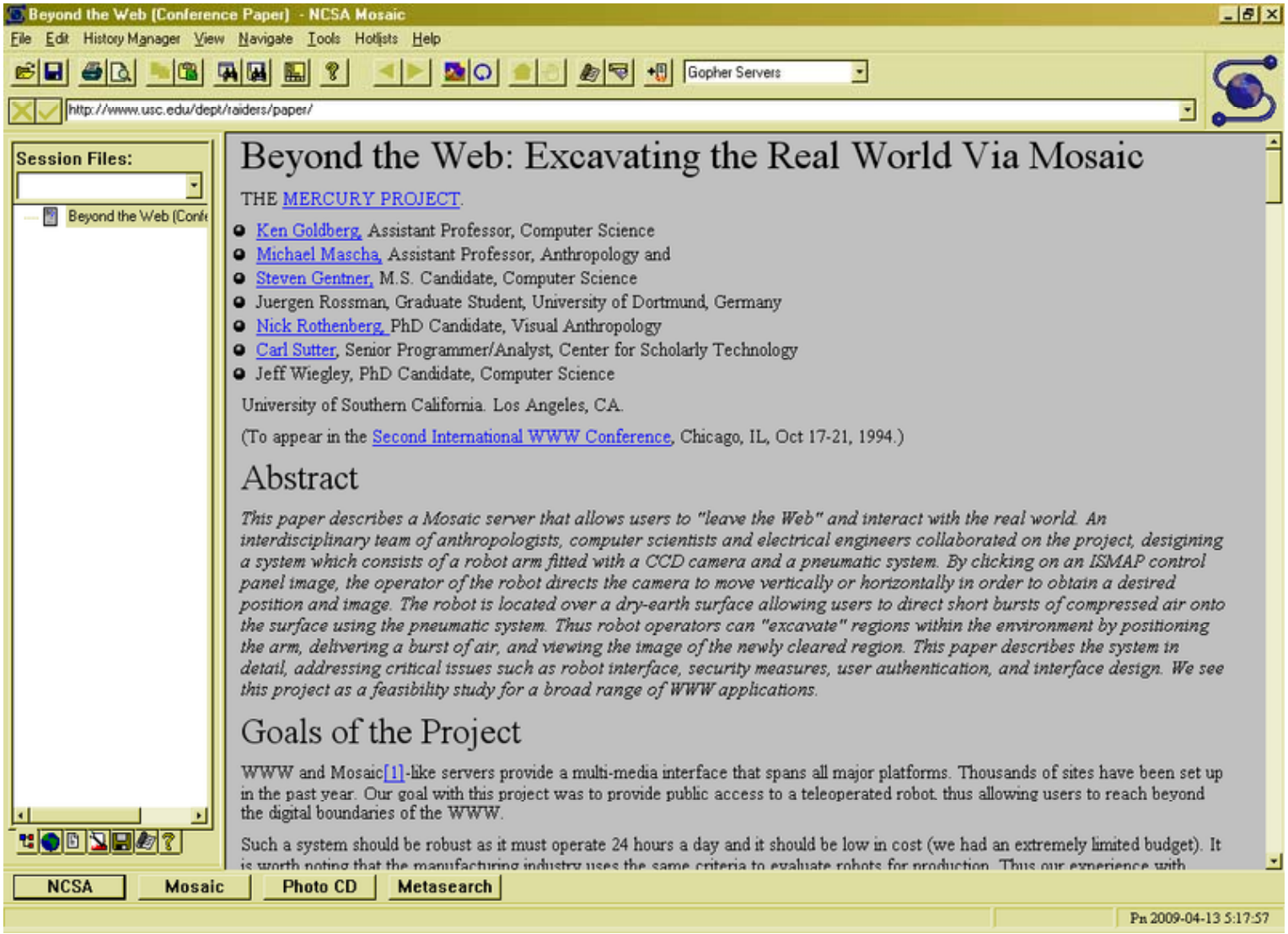
History
A summary of the history of the project.

How can I help ?
If you would like to support the web..

Getting code
Getting the code by anonymous FTP , etc.

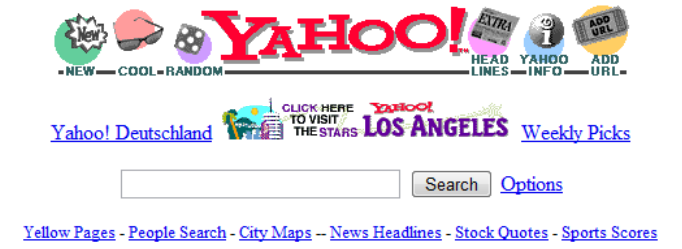
http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html

1993: First real browser – Mosaic



Brief history of the Web

- 1989:
  - HTTP, HTML
- 1990, Christmas:
  - First web server at CERN
- 1992-95: Growth
  - 1993: first Unix graphical web browser – Mosaic
  - 1994: first WWW conference, W3C is formed
- 1996-98:
  - Commercialization of the WWW
- 1999-2001:
  - Dot-com boom
- 2002:
  - Web is ubiquitous
    - Web 2.0,
    - User generated content (blogs, rss)
    - Semantic Web



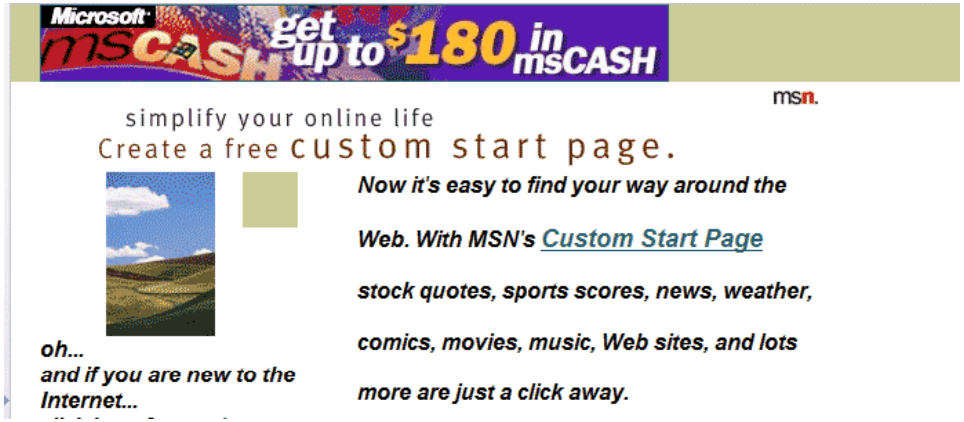
Yahoo, 1996

Amazon, 1995

Brief history of the Web

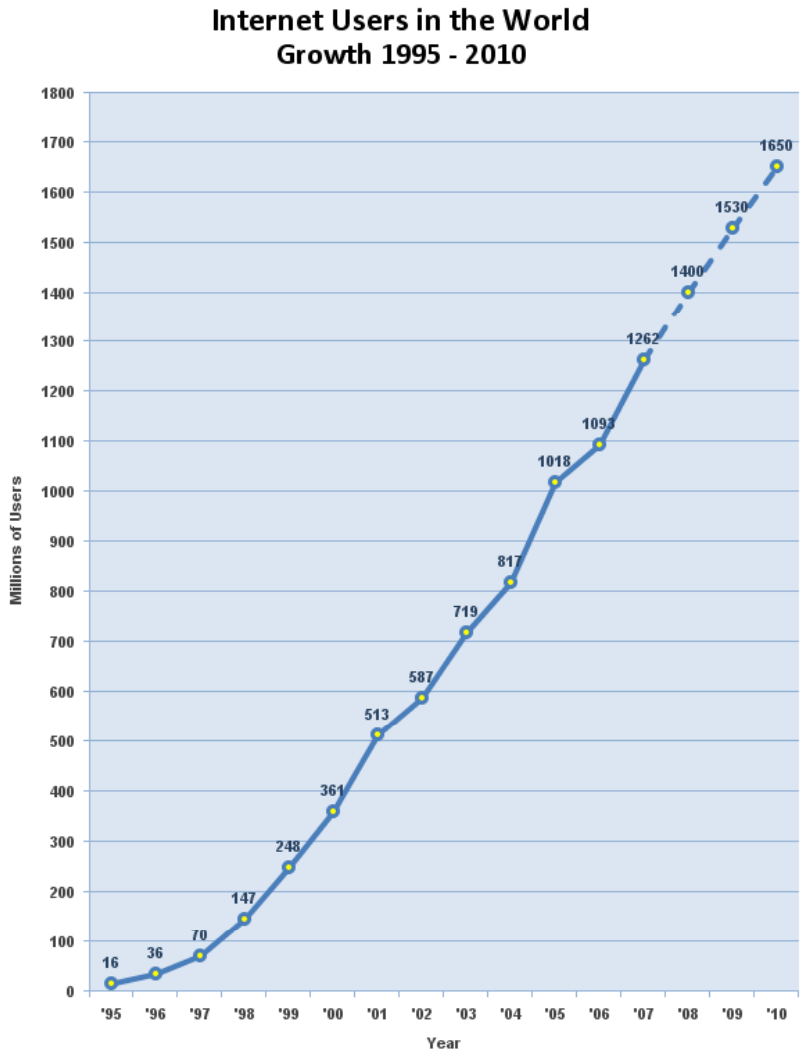
Microsoft, 1995

Google, 1998



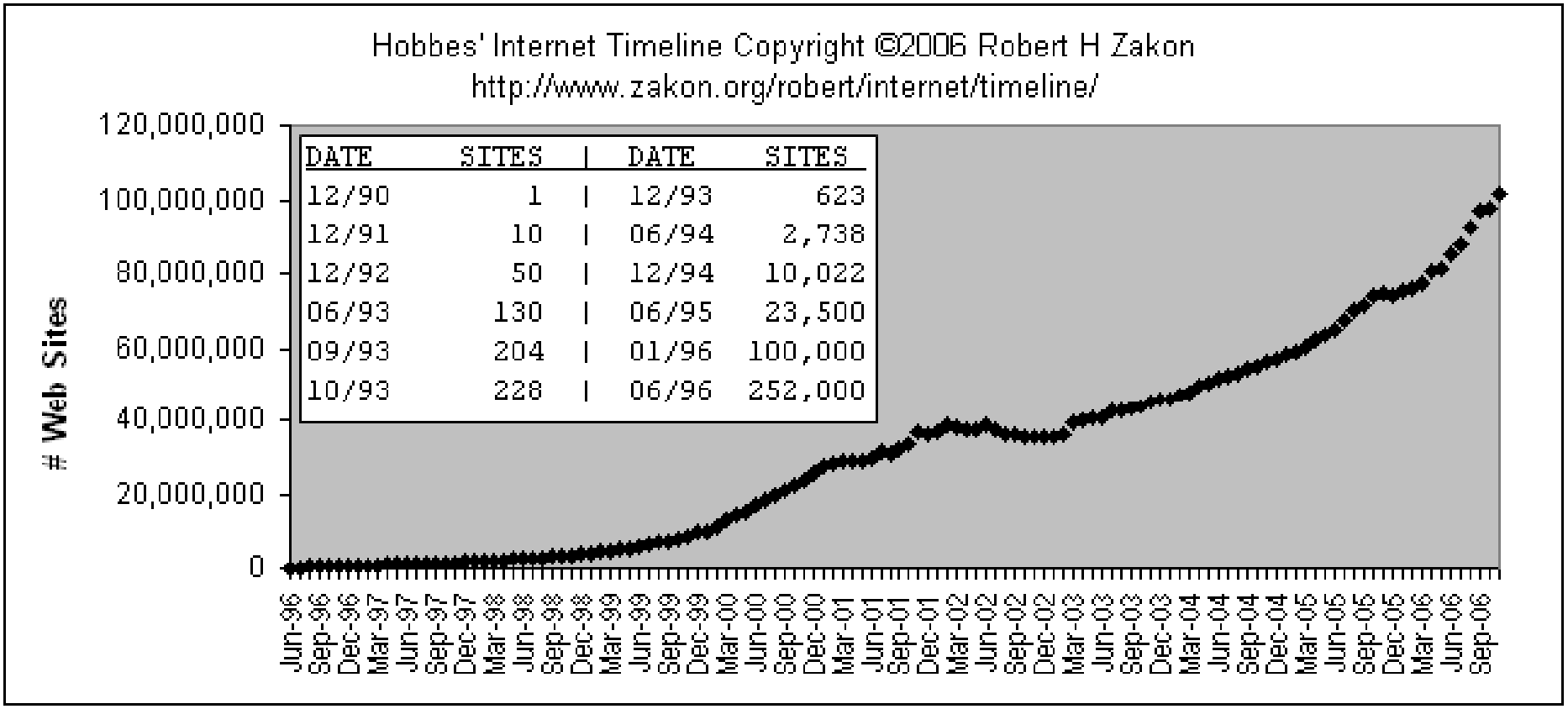
MSN, 1996

Growth of the Web: Users



Source: www.internetworldstats.com - January, 2008

Growth of the Web: Websites

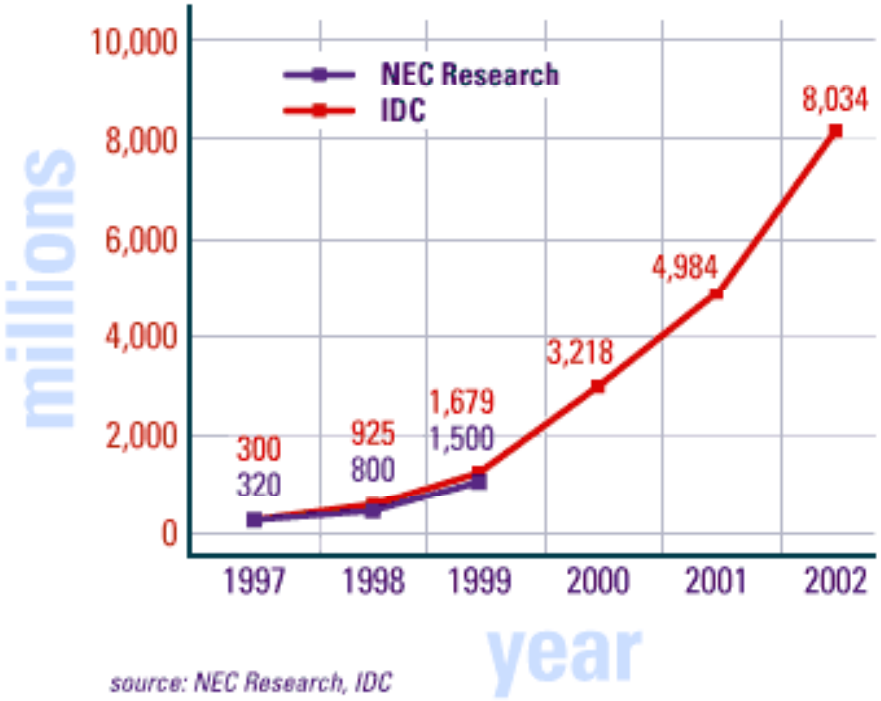


DATE	SITES	DATE	SITES
12/90	1	12/93	623
12/91	10	06/94	2,738
12/92	50	12/94	10,022
06/93	130	06/95	23,500
09/93	204	01/96	100,000
10/93	228	06/96	252,000

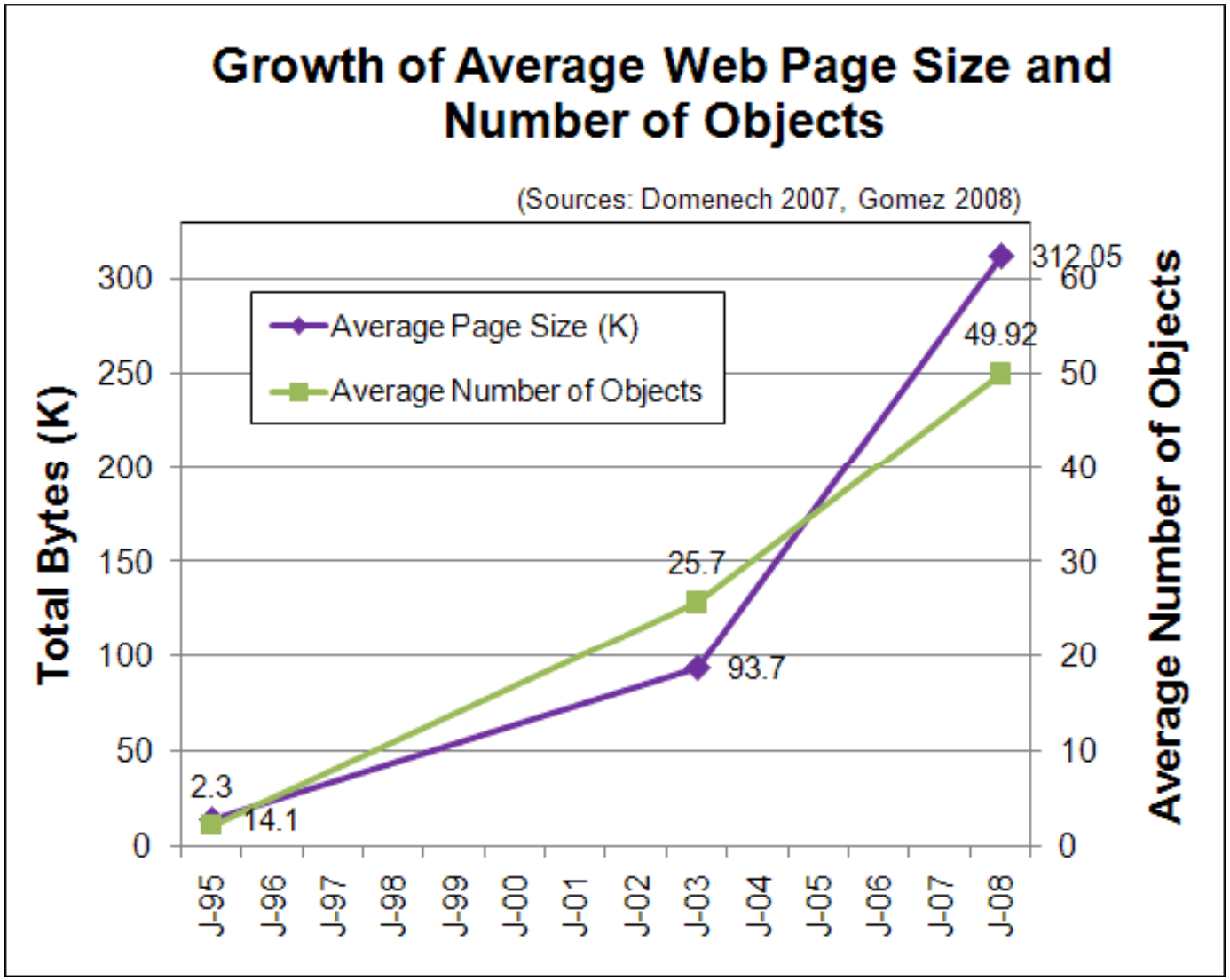
- Estimated number of websites (web servers)
- August 2009: 225 million hostnames

Growth of the Web: Pages

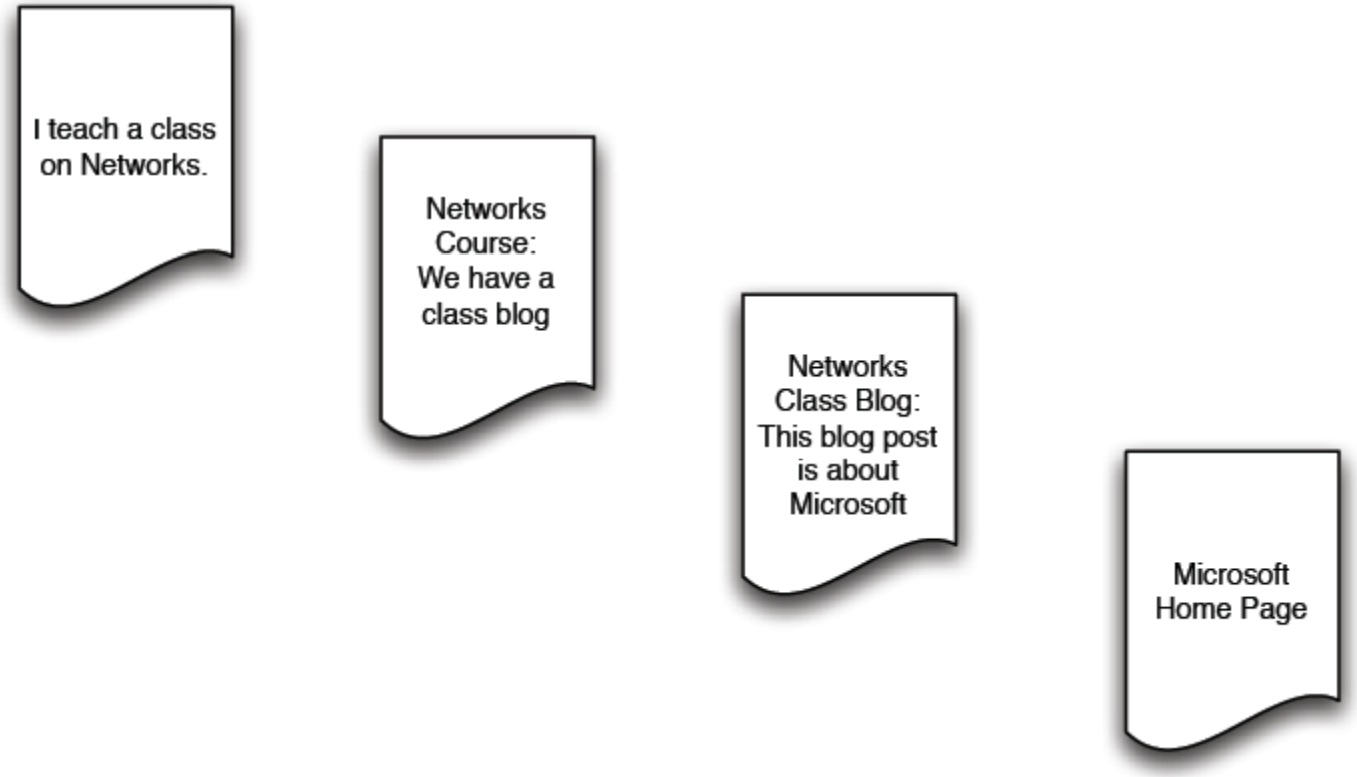
- Yahoo, 2005:
  - indexed 19.2 billion pages
- Feb 2007:
  - Estimate 29 billion pages
- Not clear:
  - How do we discover webpages?
  - Hidden web (deep web)
  - What does a page mean:
    - dynamically generated pages
    - Time stamped URLs



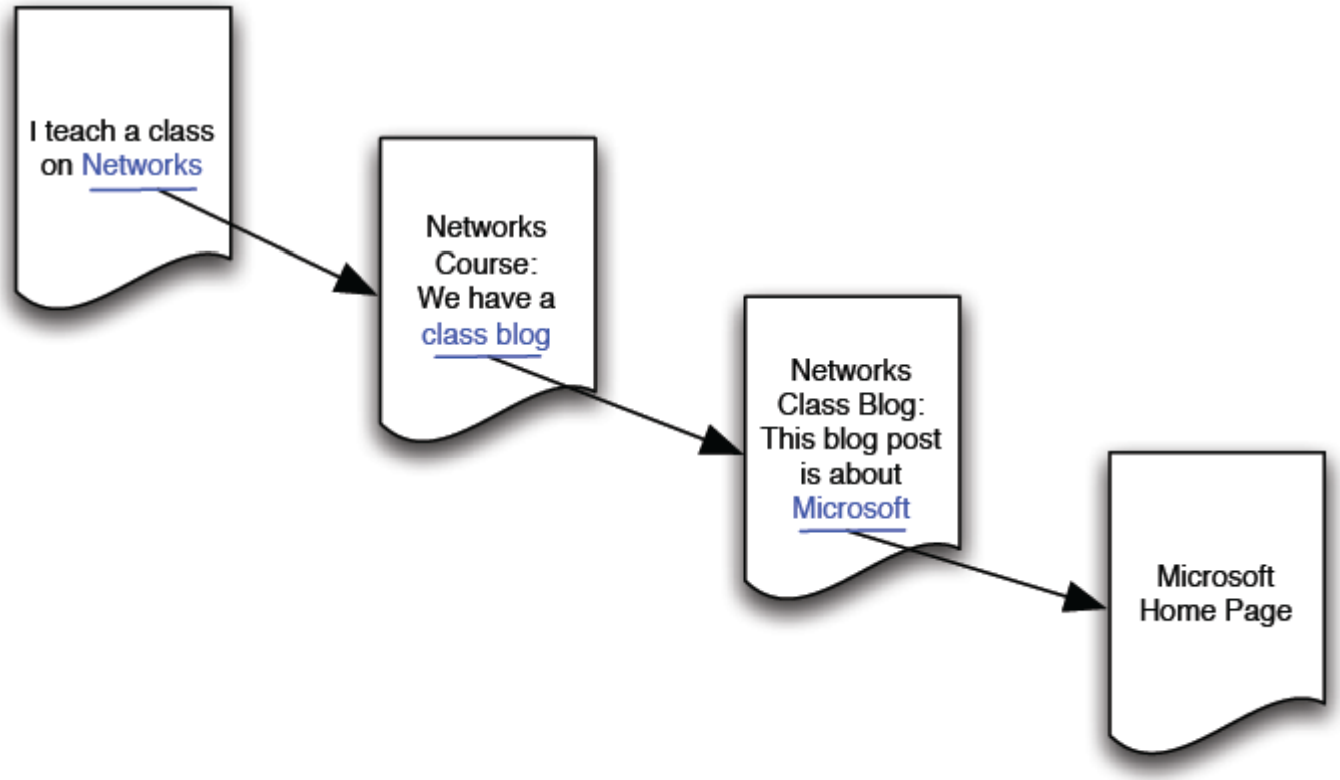
Growth of the Web: Page size



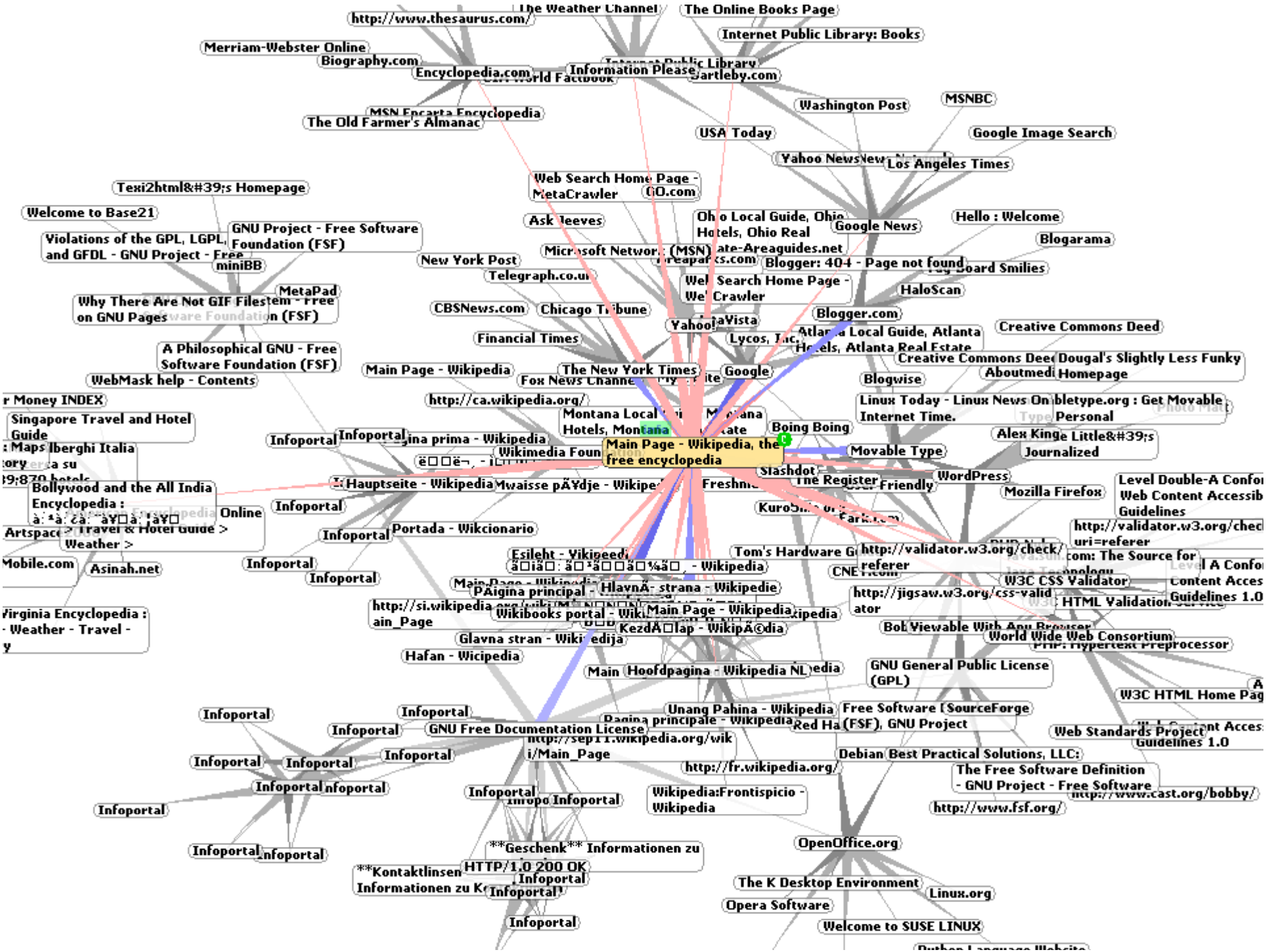
Web as a Graph



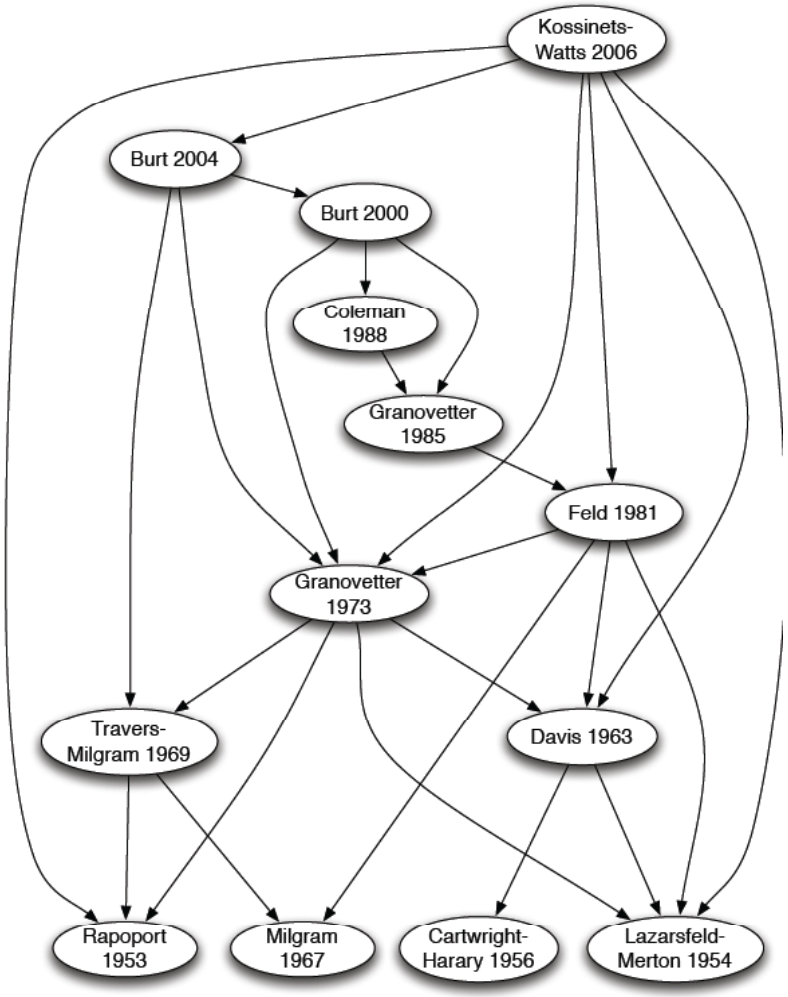
Web as a Graph



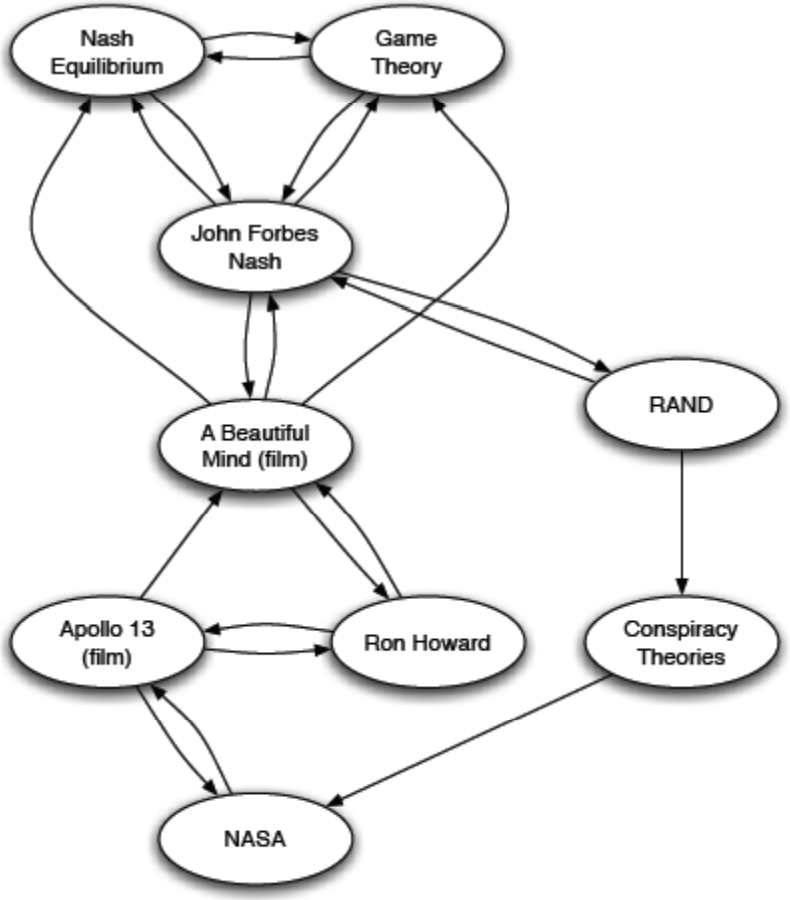
- In early days of the Web links were navigational
- Today many links are transactional


Main Page - Wikipedia, the free encyclopedia
http://www.thesaurus.com/
The Weather Channel
The Online Books Page
Internet Public Library: Books
Merriam-Webster Online
Biography.com
Encyclopedia.com
Information Please
Bartleby.com
MSN Encarta Encyclopedia
The Old Farmer's Almanac
Washington Post
MSNBC
USA Today
Google Image Search
Yahoo News
Los Angeles Times
Texi2html's Homepage
Welcome to Base21
Violations of the GPL, LGPL, and GFDL - GNU Project - Free
GNU Project - Free Software Foundation (FSF)
miniBB
Web Search Home Page - MetaCrawler
Ask Jeeves
Google News
Hello : Welcome
Blogarama
New York Post
Telegraph.co.uk
Blogger: 404 - Page not found
HaloScan
MetaPad
Why There Are Not GIF Files on GNU Pages
CBSNews.com
Chicago Tribune
Blogger.com
Financial Times
Creative Commons Deed
A Philosophical GNU - Free Software Foundation (FSF)
WebMask help - Contents
Main Page - Wikipedia
The New York Times
Google
Fox News Channel
Blogwise
Dougal's Slightly Less Funky Homepage
Aboutmedi
Singapore Travel and Hotel Guide
http://ca.wikipedia.org/
Linux Today - Linux News On Internet Time.
Personal
Montana Local Hotels, Montana
Infoportal
Boing Boing
Alex King
Little's Journalized
Wikimedia Foun
Movable Type
Hauptseite - Wikipedia
Slashdot
The Register
WordPress
Mozilla Firefox
Bollywood and the All India Encyclopedia :
Travel & Hotel Guide >
Weather >
Portada - Wikcionario
Asinah.net
Esileht - Wikipeedi
Tom's Hardware
http://validator.w3.org/check/referer
W3C CSS Validator
W3C HTML Validation
http://jigsaw.w3.org/css-validator
Wikibooks portal - Wikipedia
Main Page - Wikipedia
KezdÃ□lap - WikipÃ©dia
Glavna stran - Wikipedija
Hafan - Wicipedia
Hoofdpagina - Wikipedia
GNU General Public License (GPL)
World Wide Web Consortium
W3C HTML Home Page
Unang Pahina - Wikipedia
Free Software (SourceForge
Red Ha
(FSF), GNU Project
GNU Free Documentation License
Web Standards Project
Debian
Best Practical Solutions, LLC:
http://fr.wikipedia.org/
The Free Software Definition - GNU Project - Free Software
http://www.fsf.org/
Wikipedia:Frontispicio - Wikipedia
OpenOffice.org
Geschenk Informationen zu
HTTP/1.0 200 OK
**Kontaktlinsen
The K Desktop Environment
Linux.org
Opera Software
Welcome to SUSE LINUX
Infoportal

Other information networks



Citations

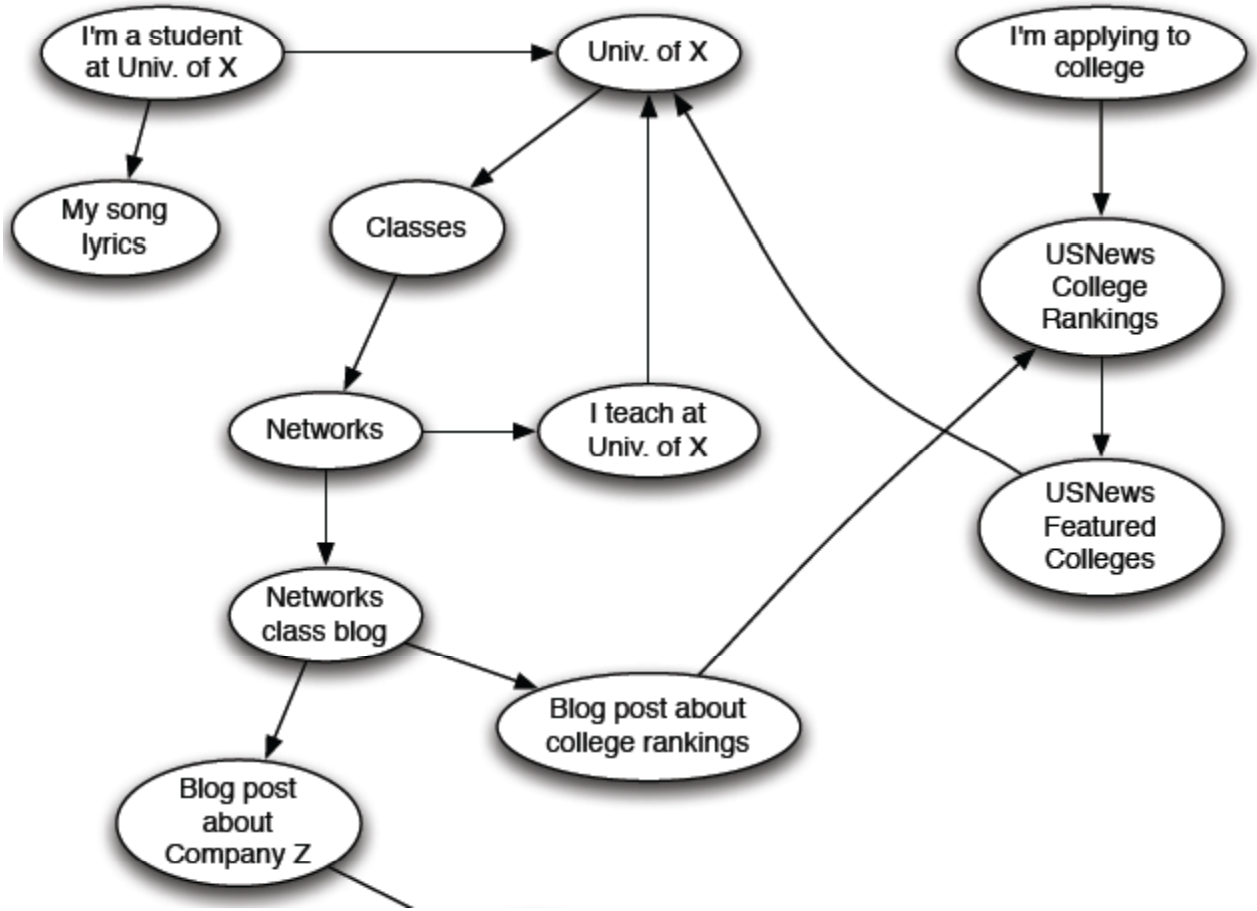


References in enciclopedia

How does the Web look like?

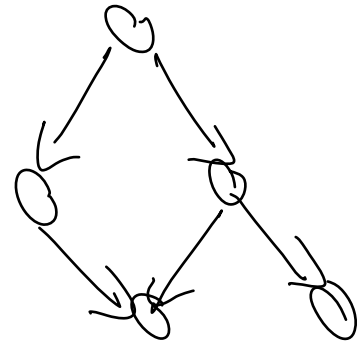
- How is information on the Web organized?
- How is the Web linked?
- What is the “map” of the Web?

Web as a directed graph

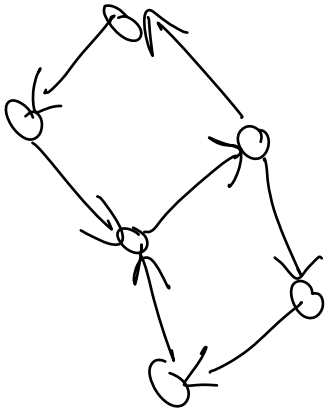


Directed graphs

- Two types of directed graphs:
  - DAG – directed acyclic graph:
    - Has no cycles: if u can reach v, then v can not reach u



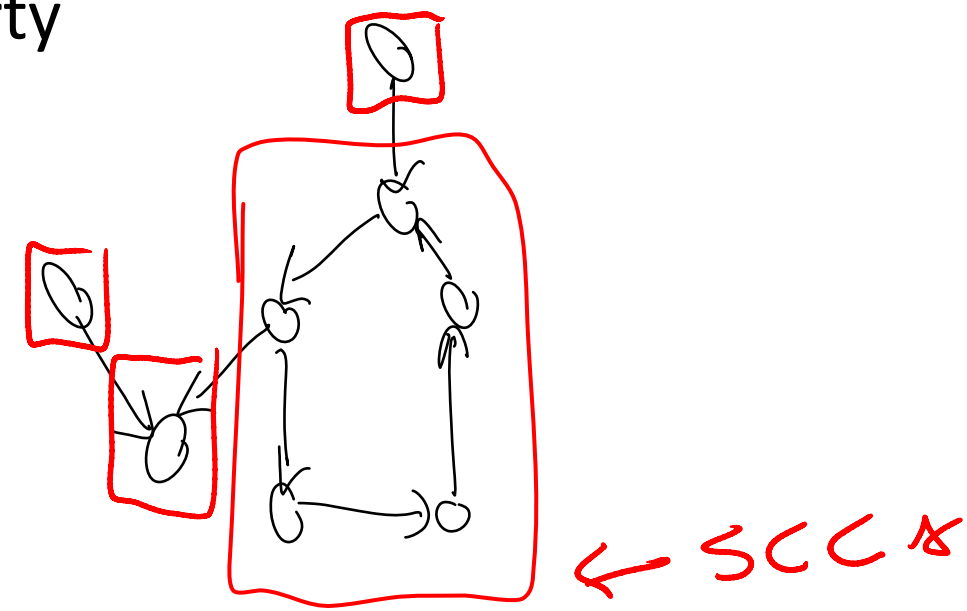
  - Strongly connected:
    - Any node can reach any node via a directed path



- Any directed graph can be expressed in terms of these two types

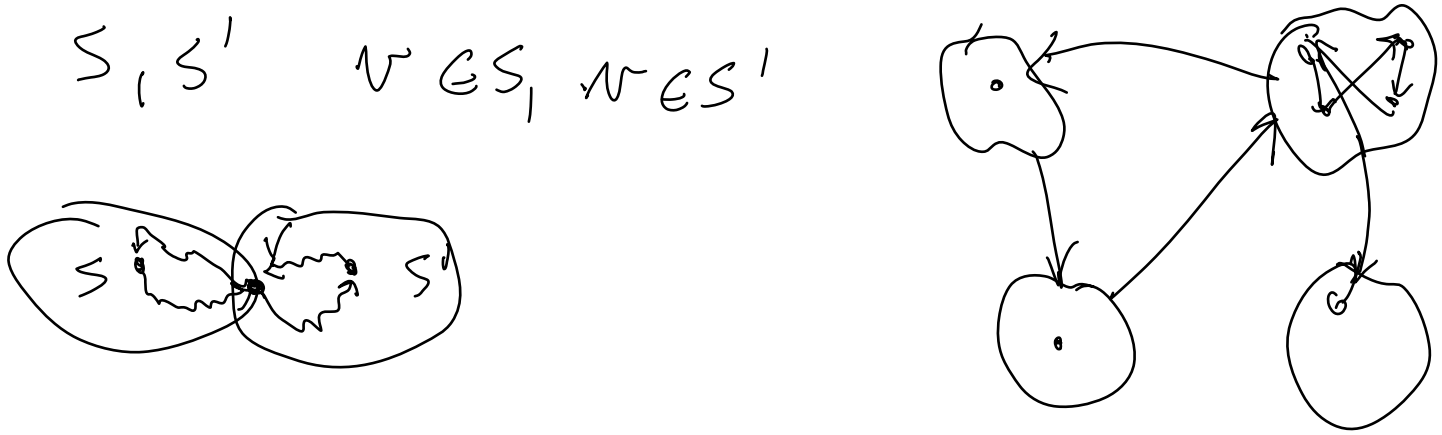
Strongly connected component

- Strongly connected component (SCC) is a set of nodes S so that:
  - Every pair of nodes in S can reach each other
  - There is no larger set containing S with this property



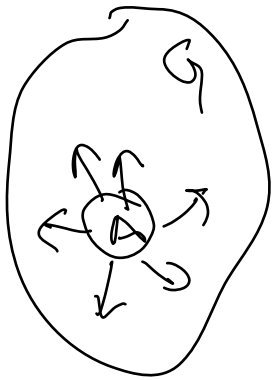
Strongly connected component

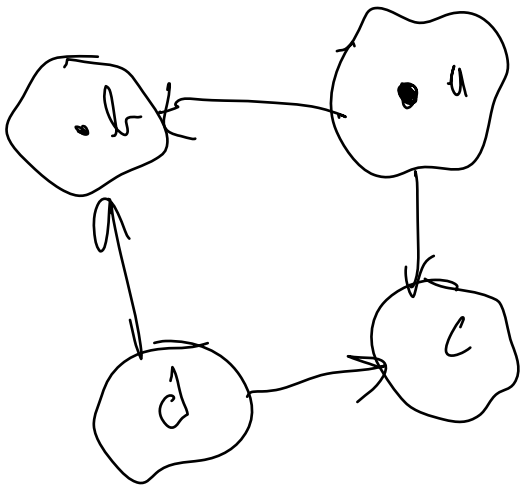
- Fact: Every directed graph is a DAG on its SCCs.
  - SCCs partition the nodes of G (each node in exactly one SCC)
  - If we build a graph G' whose nodes are SCCs, and edge between nodes G' if there is an edge between corresponding SCCs in G, then G' is a DAG



Example

- Given a directed graph G
- Find the smallest set of node A so that every node in G is reachable from at least one node in A.

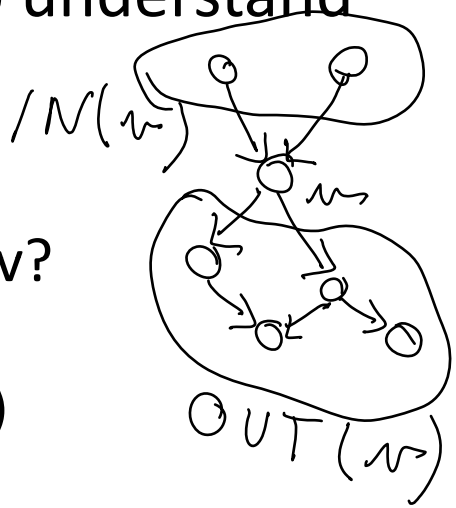




$u \in A$

Graph structure of the Web

- Take a large snapshot of the web and try to understand how it's SCCs "fit" as a DAG.



- Computational issue:
  - Say want to find SCC containing specific node v?
  - Observation:
    - Out(v) … nodes that can be reached from v (BFS out)
    - SCC containing v is:

$= OUT(v) \cap IN(v)$

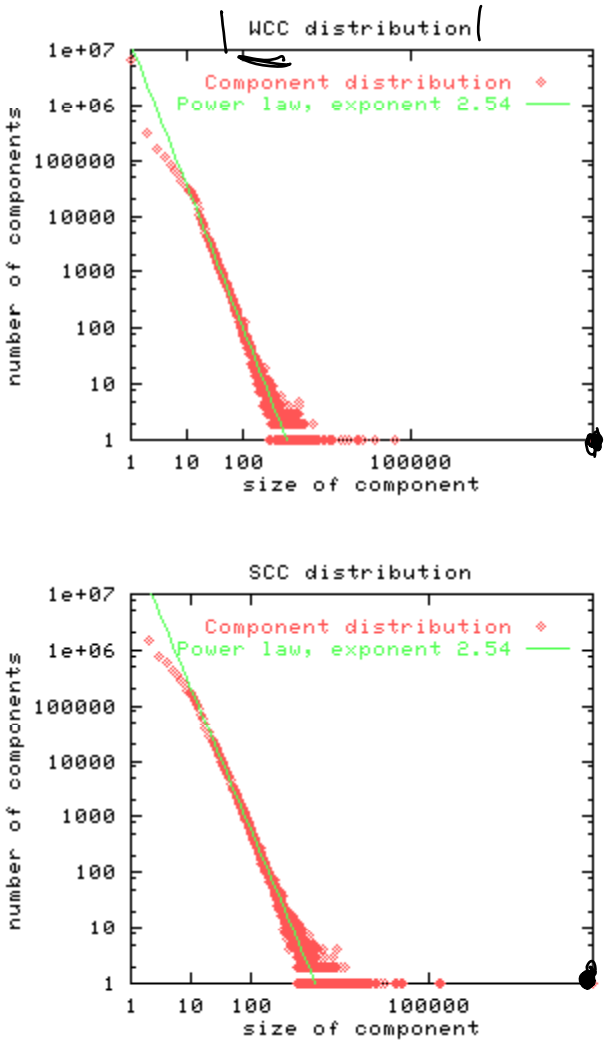
$= OUT(v, G) \cap OUT(v, \bar{G})$

$\bar{G}$ … G with edge directions flipped

[Broder et al., 'oo]

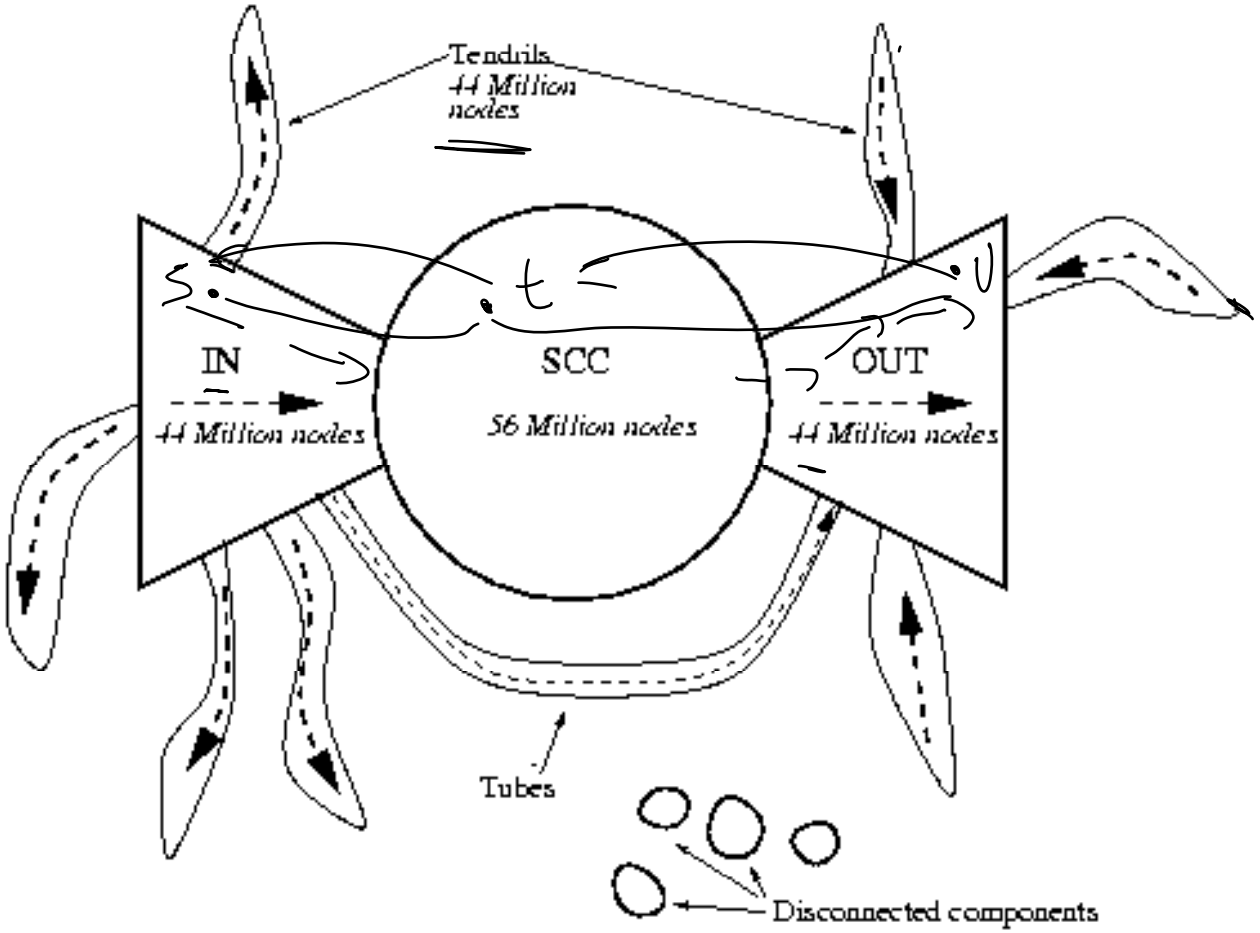
Graph structure of the Web

- There is a giant SCC
- There will not be 2 giant SCCs:
  - Just takes 1 page from each to link to one in other – if the components have millions of pages the likelihood of this is large
- Broder et al., 2000:
  - Weakly connected component: 90% of the nodes



[Broder et al., 'oo]

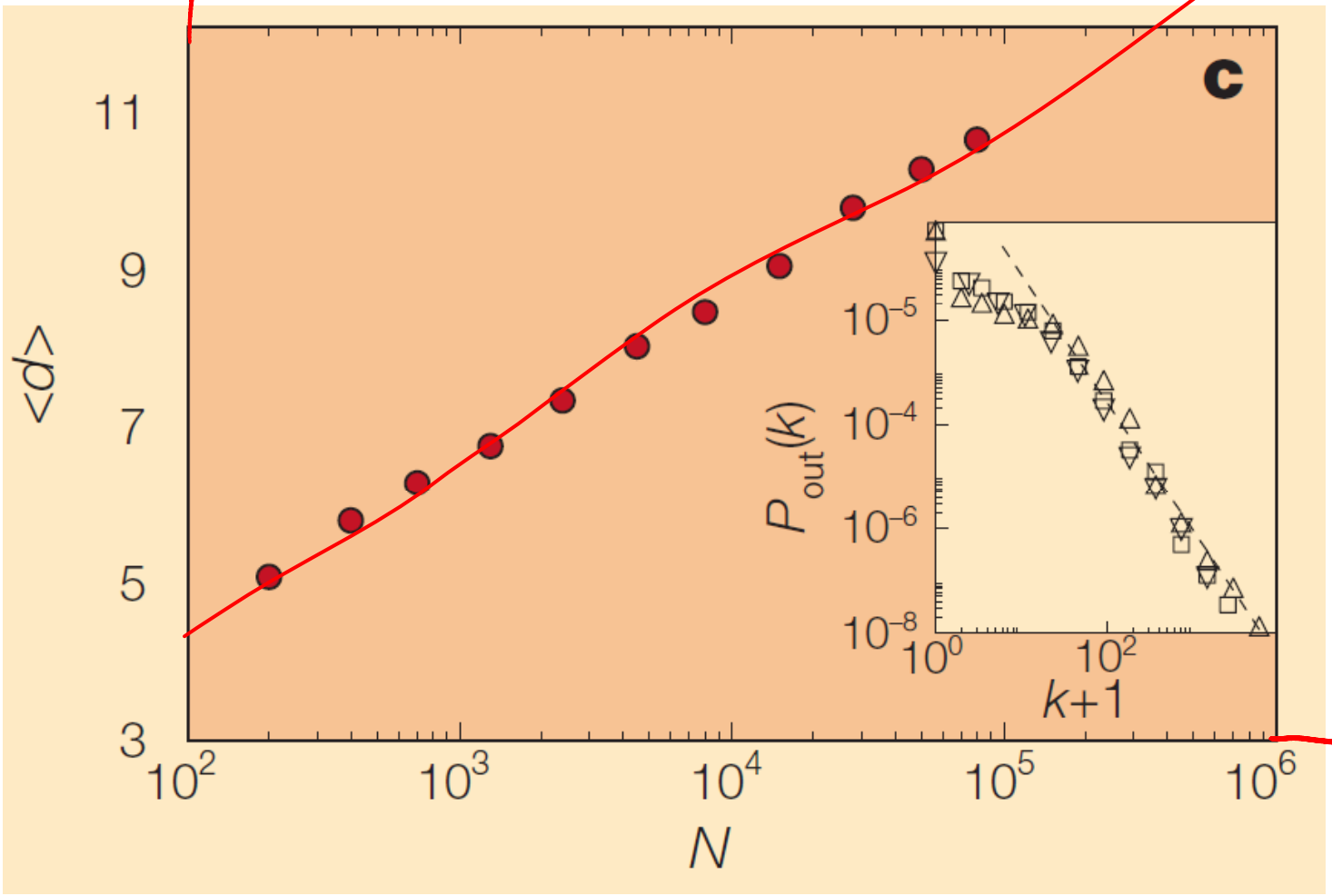
Bow-tie structure of the Web



- 250 million webpages, 1.5 billion links [Altavista]

Diameter of the Web

[Albert et al., '99]

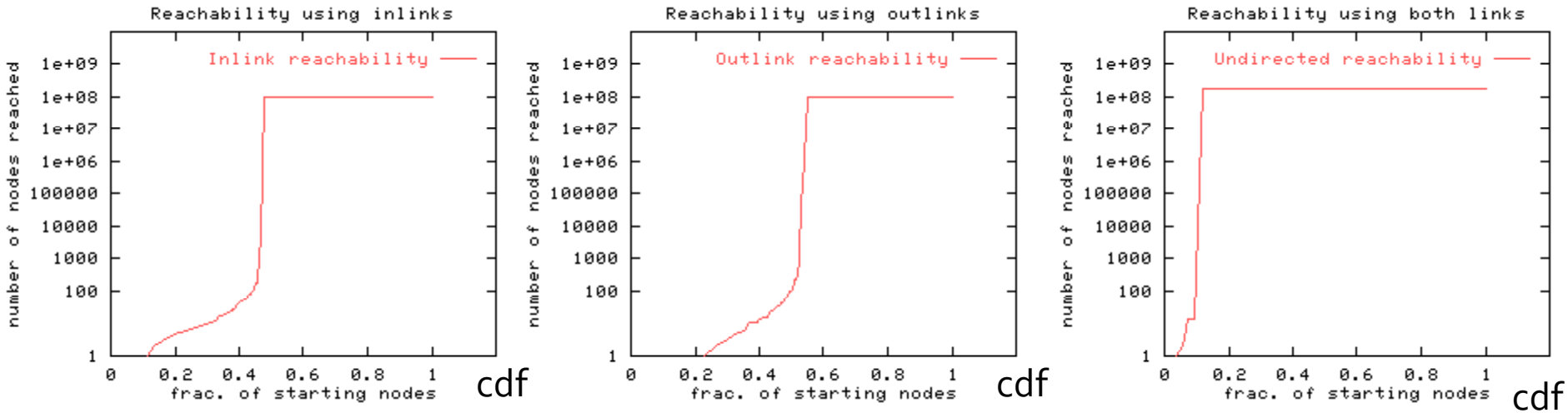


- Diameter (average directed shortest path length) is 19 (in 1999)

[Broder et al., 'oo]

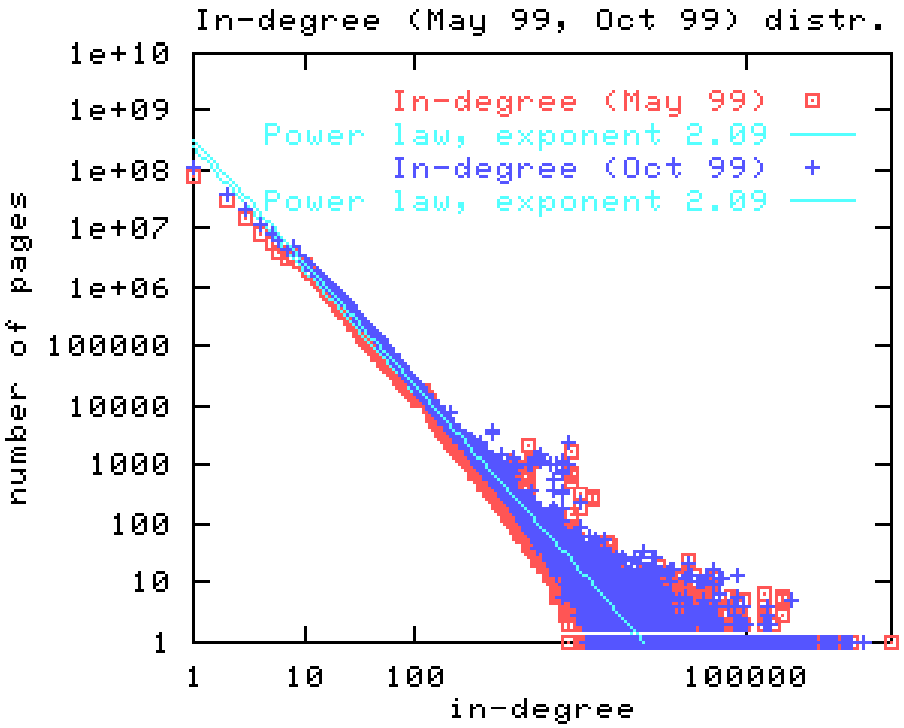
Diameter of the Web

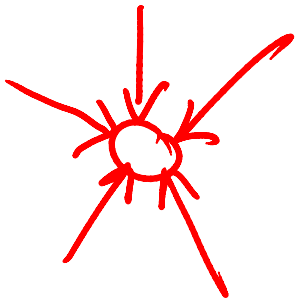
- Average distance: 75% of time there is no directed path
  - Follow in-links (directed): 16.12
  - Follow out-links (directed): 16.18
  - Undirected: 6.83
- Diameter of SCC (directed): At least 28

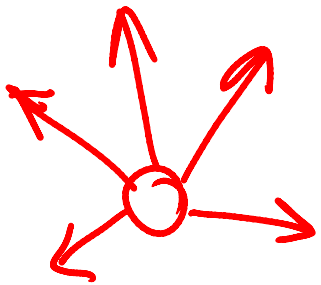


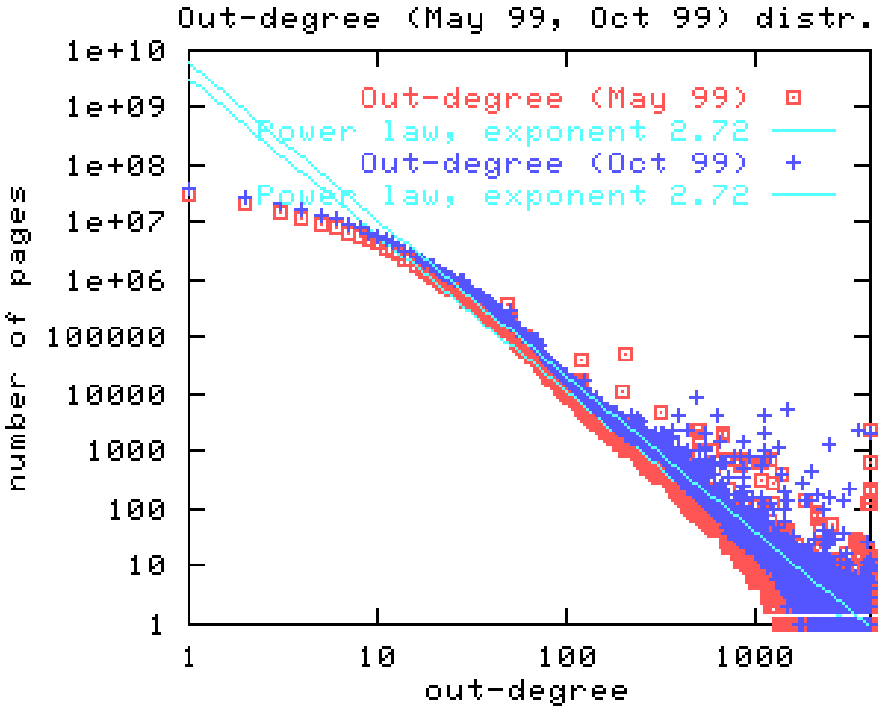
[Broder et al., 'oo]

Degree distribution on the Web





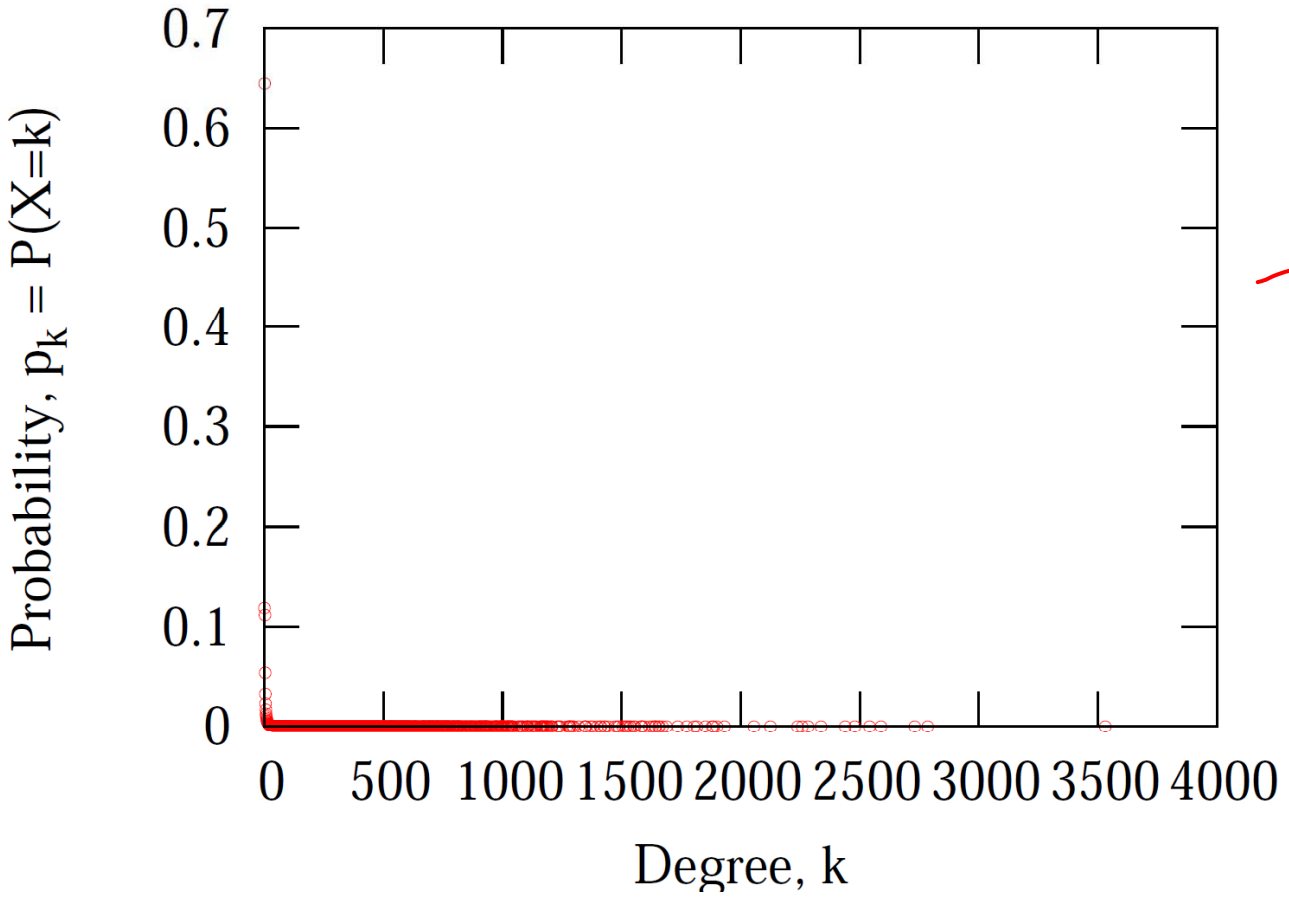


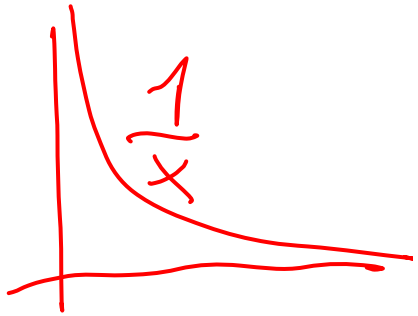


[Leskovec et al. KDD '08]

Degrees in real networks

- Take real network plot a histogram of p_k vs. k



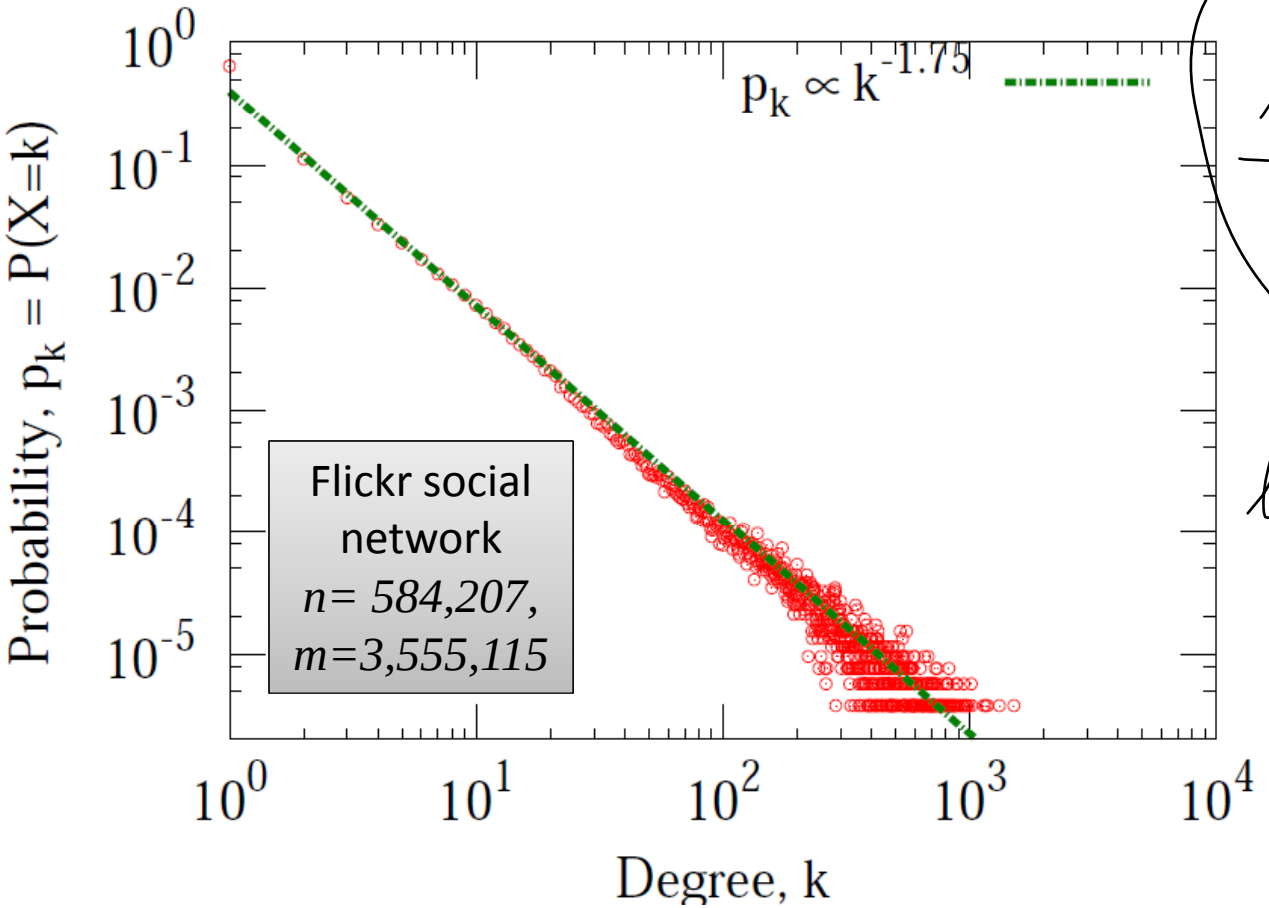


Flickr social network
n= 584,207,
m=3,555,115

[Leskovec et al. KDD '08]

Degrees in real networks (2)

- Plot the same data on *log-log* axis:



$f(k) \sim c \cdot k^{-\alpha}$

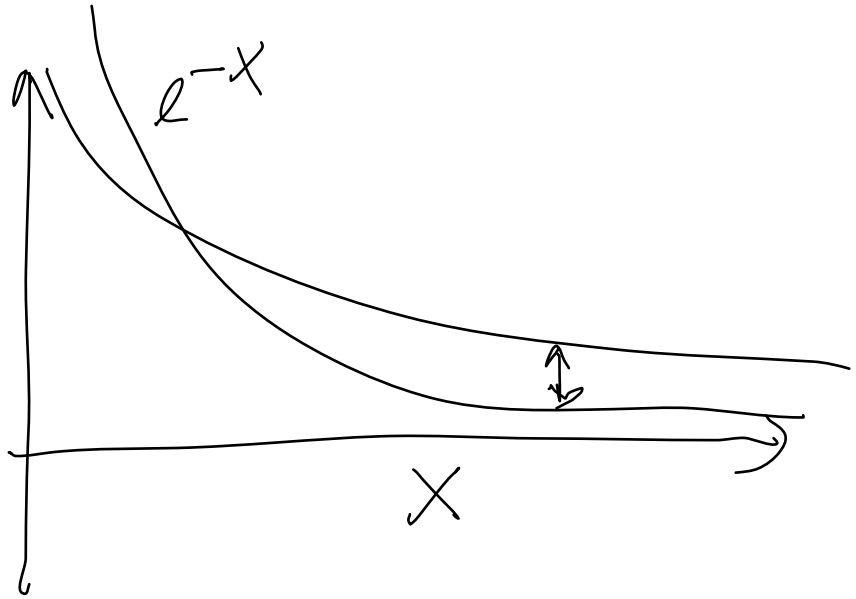
$\log f(k) = \log c + -\alpha \log k$

Heavy tails

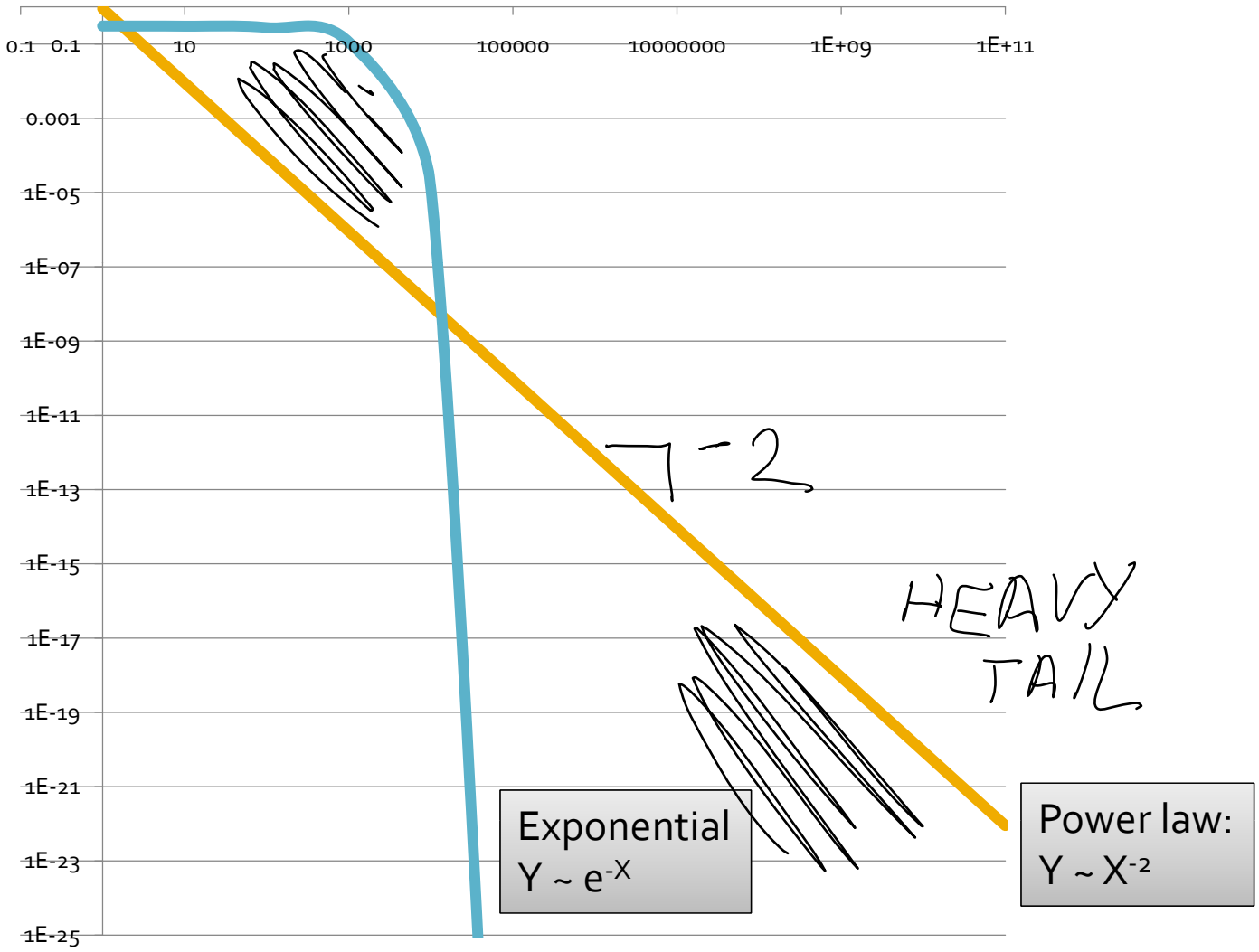
- Degrees are heavily skewed:

  Distribution is heavy tailed:

$$\lim_{x \to \infty} \frac{Pr(X > x)}{e^{-\epsilon x}} = \infty$$



Exponential tail vs. Power-law tail



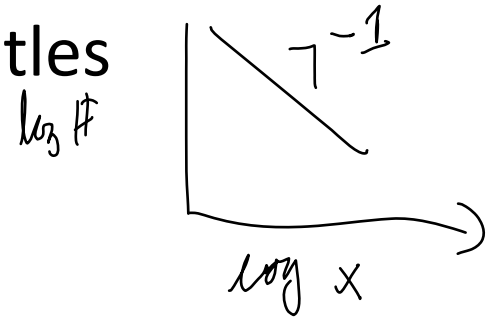
Heavy tails

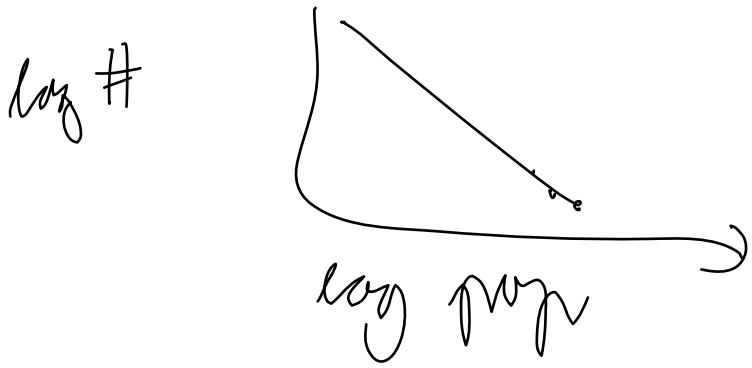
- Various names, kinds and forms:
  - Long tail, Heavy tail, Zipf's law, Pareto's law, 80-20 rule

power law	$x^{-\alpha}$
power law with cutoff	$x^{-\alpha}\mathrm{e}^{-\lambda x}$
stretched exponential	$x^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$
log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$

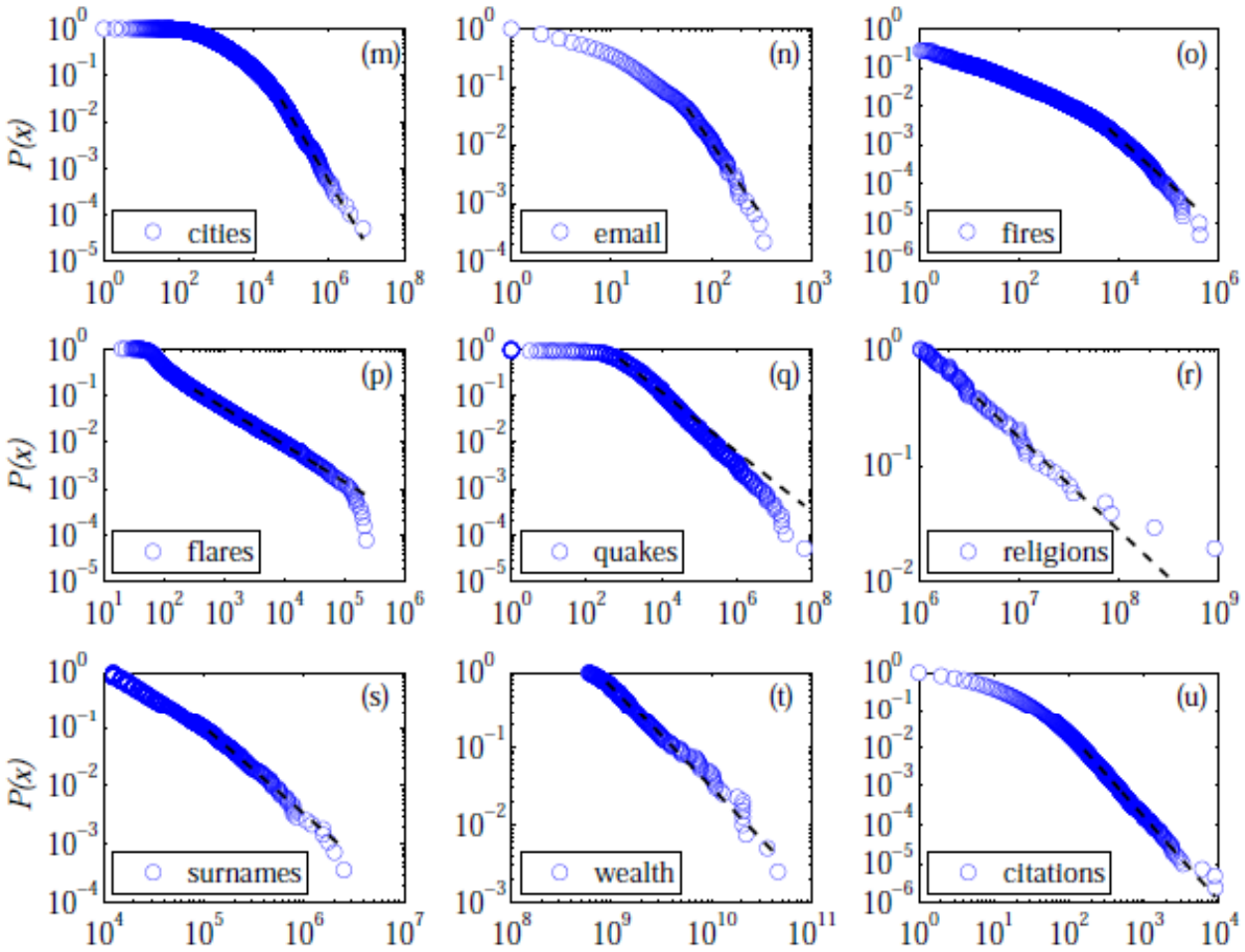
Power-laws are everywhere

- In social systems – lots of power laws:
  - Pareto, 1897 – Wealth distribution
  - Lotka 1926 – Scientific output
  - Yule 1920s -- lots of species of beetles
  - Zipf 1940s – word frequency
  - Simon 1950s – city populations



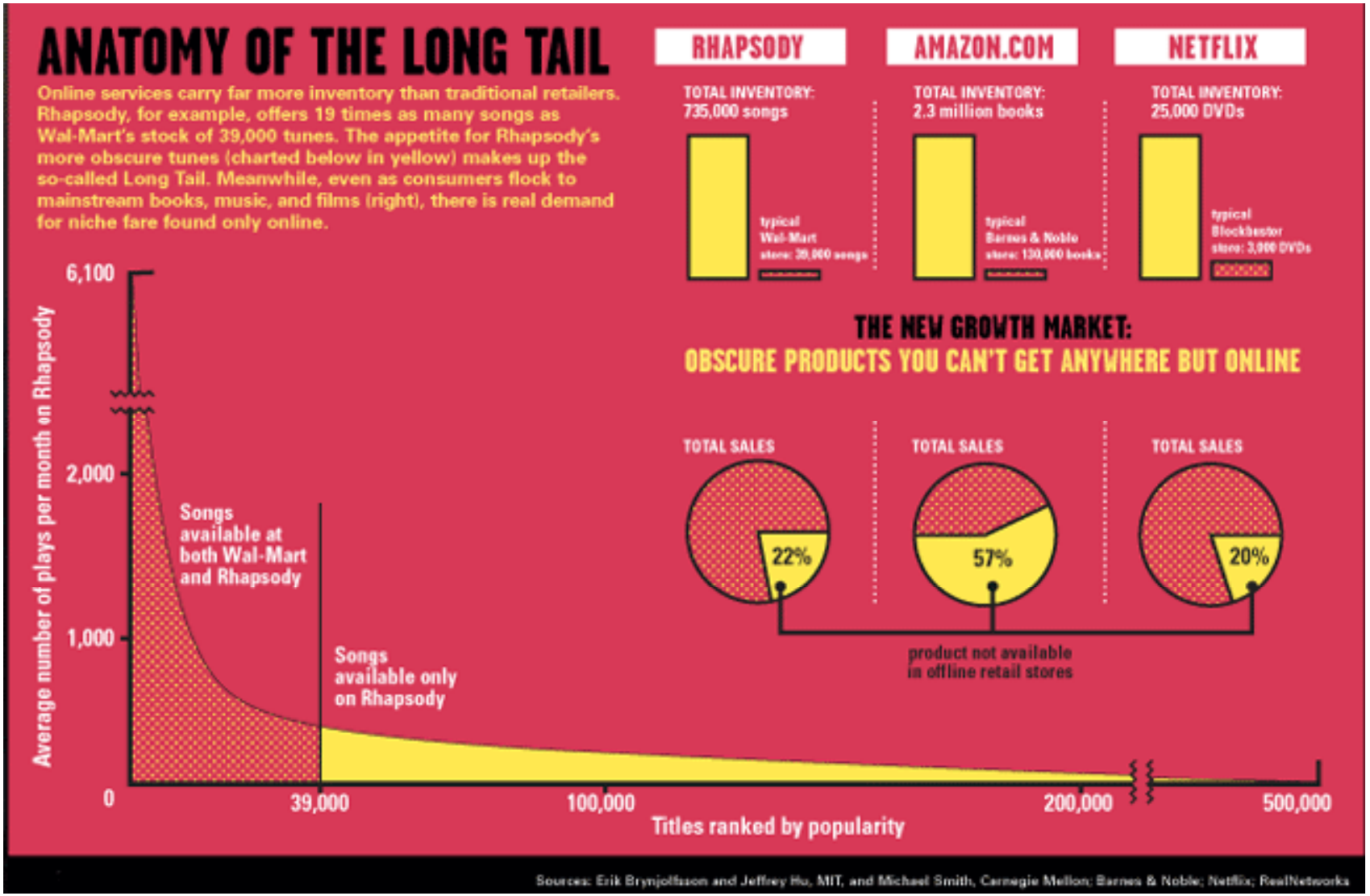


Power-laws are everywhere



Many other quantities follow heavy-tailed distributions

The long tail



Not everyone likes power-laws ☺

CMU students protesting at the G20 meeting in Pittsburgh in Sept 2009