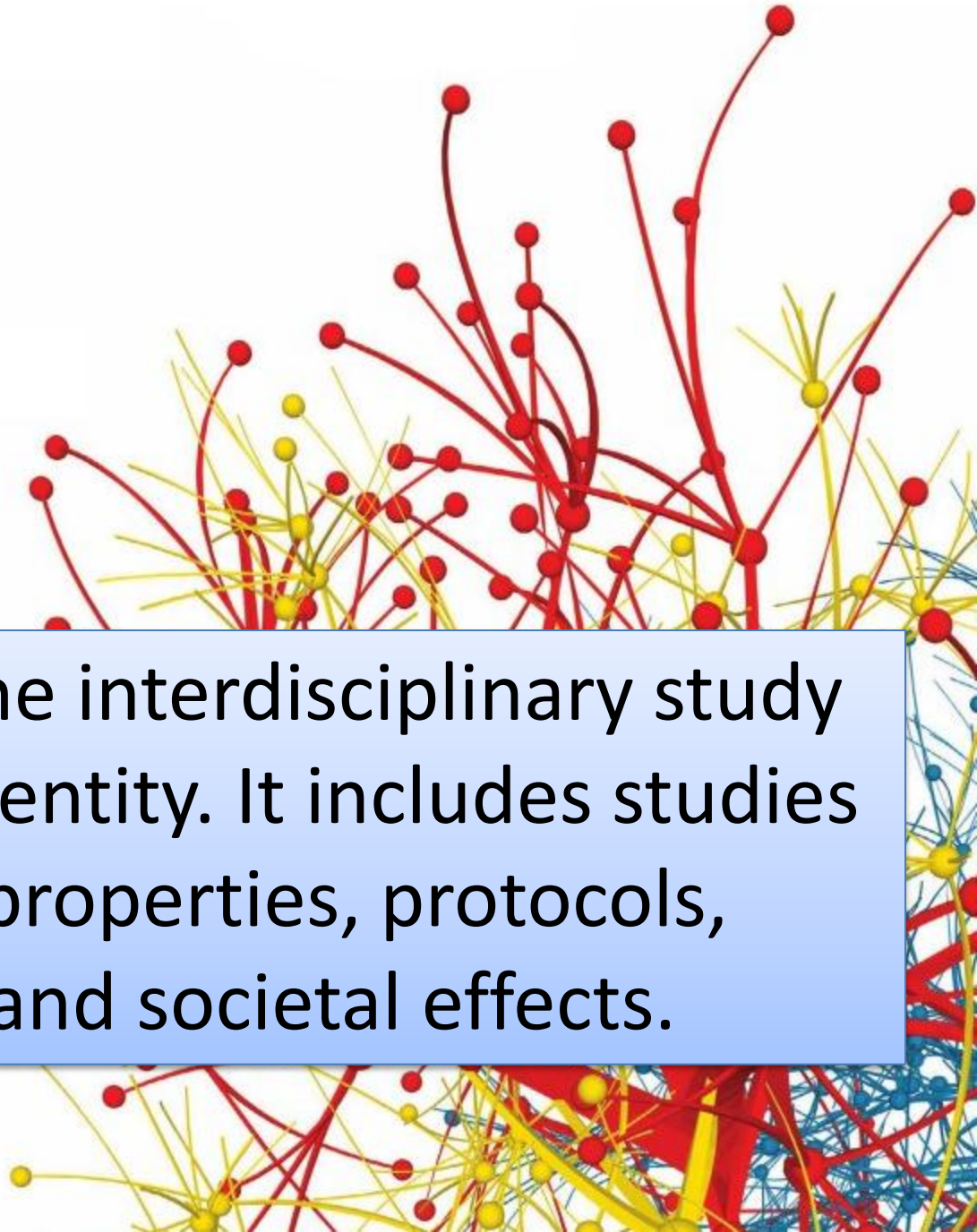
Introduction to Web Science

Intro to Web Science

IB Computer Science (Higher)

2014

What is Web Science?

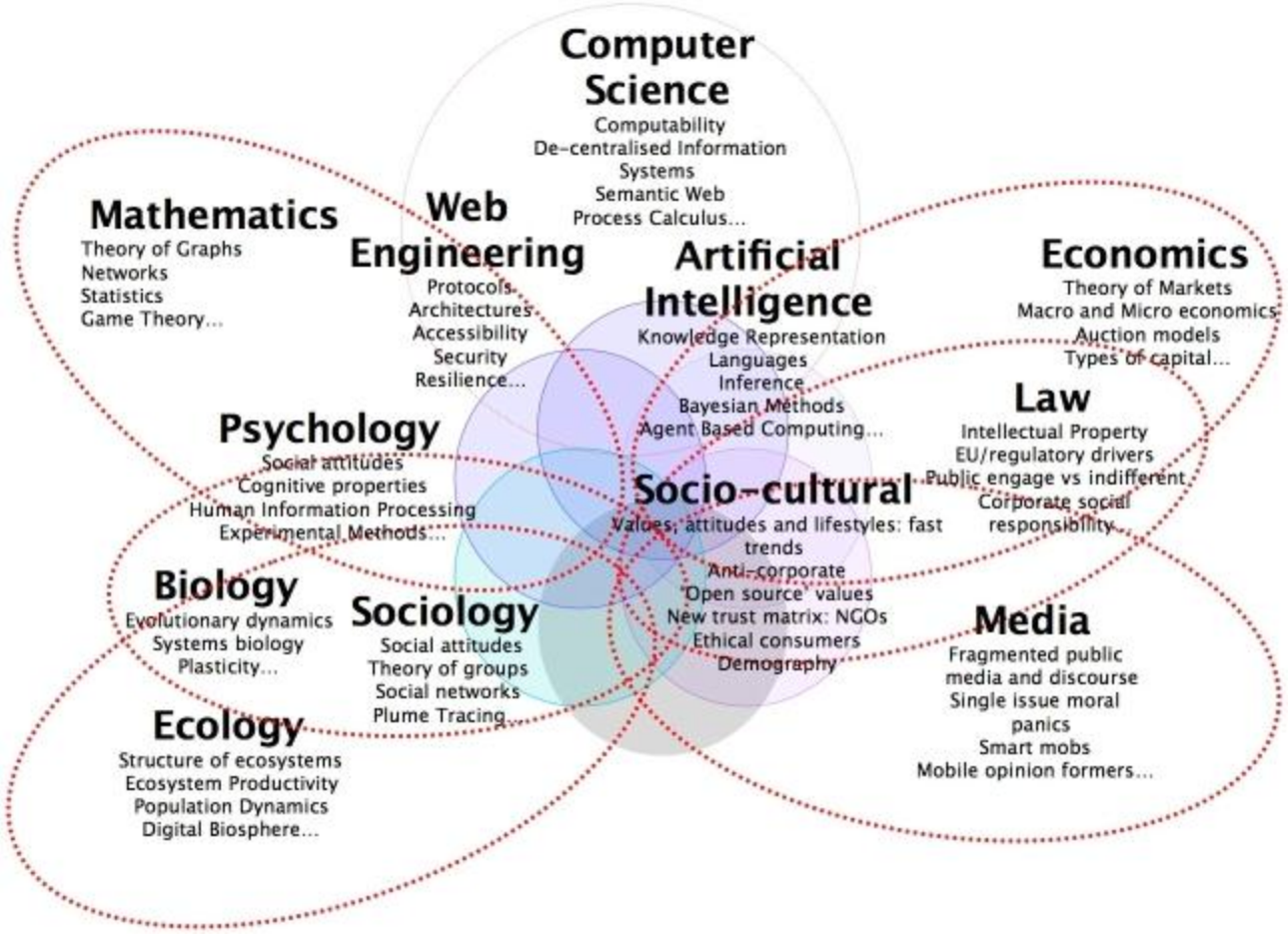


Web Science is the interdisciplinary study of the Web as an entity. It includes studies of the Web's properties, protocols, algorithms, and societal effects.

http://mags.acm.org/communications/200807/#pg1

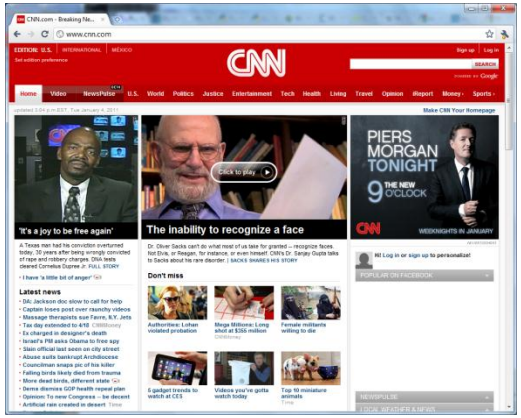
“Given the breadth of the Web and its inherently multi-user (social) nature, its **science is necessarily interdisciplinary**, involving at least mathematics, CS, artificial intelligence, sociology, psychology, biology, and economics. We invite computer scientists to **expand the discipline** by addressing the challenges following from the widespread adoption of the Web and its profound influence on social structures, political systems, commercial organisations, and educational institutions.”

Web Science is Interdisciplinary



O'Hara and Hall, *Web Science*, ALT Online Newsletter, May 6, 2008

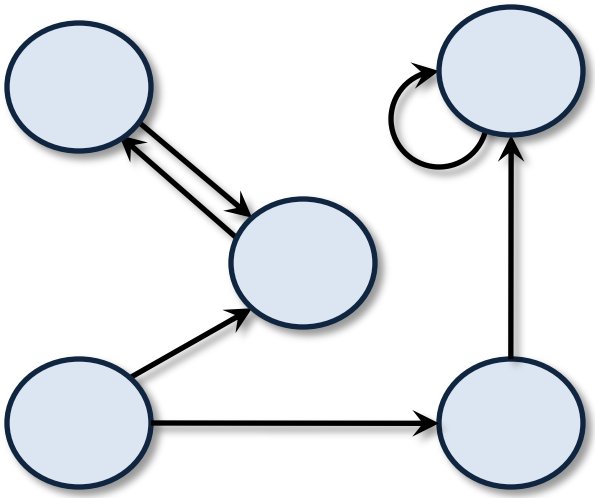
How is the Web structured?



link

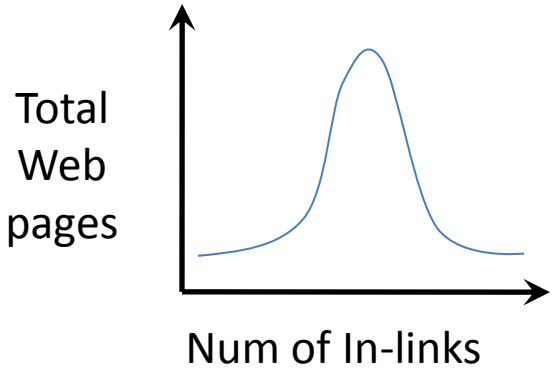
Graph Theory: Pages are nodes &
links are directed edges

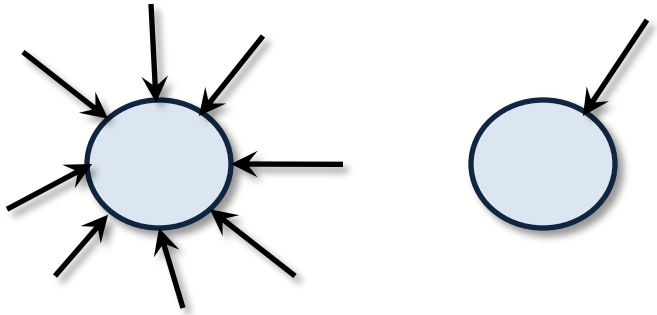
Web Graph



Random Graph

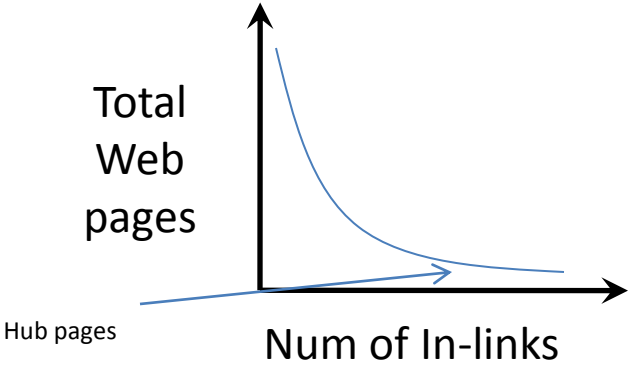
Normal/Gaussian Distribution



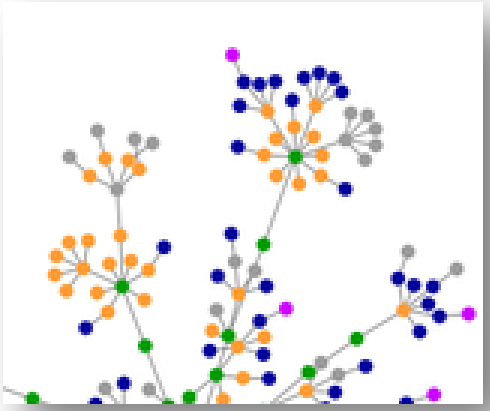


Typical Web Graph

Power-law Distribution

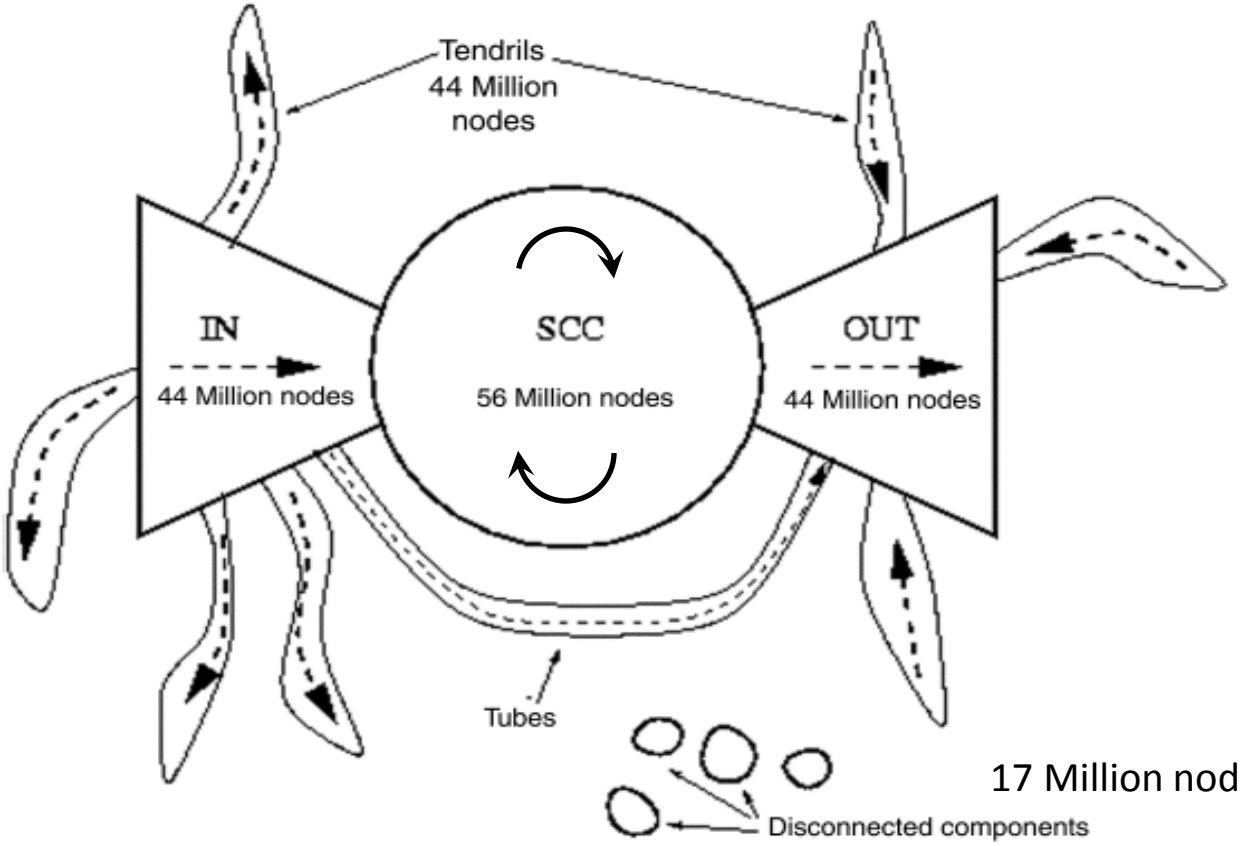


Small World Network



- Six degrees of separation
- Most pages are not neighbours but most pages can be reached from others by a small number of hops
- Many hubs- pages with many inlinks
- Robust for random node deletions
- Other examples: road maps, networks of brain neurons, voter networks, and social networks

Bow-Tie Structure of the Web



Broder et. al (Graph Structure of the Web, 2000)
Examined a large web graph (200M pages, 1.5B links)

Bow-Tie Structure

- 75% of pages do not have a direct path from one page to another
- Ave distance is **16 clicks** when path exists and **7 clicks** when undirected path exists
- Diameter of SCC is at least 28 (max shortest distance between any two nodes)
- Diameter of entire Web is at least 500 (most distant node in IN to OUT)

Broder et al., Graph Structure of the Web, 2000

Web Structure's Implications

- If we want to discover every web page on the Web, it's impossible since there are many pages that aren't linked to
- Finding popular pages is easy, but finding pages with few in-links (the long tail) is more difficult
- How do we know when new pages are added to the Web or removed?
- Incoming links could tell us something about the "importance" of a page when searching the Web for information (e.g., PageRank)
- Link structure of the Web can be artificially manipulated

How large is the Web?

Web Crawler

Web crawlers are used to fetch a page, place all the page's links in a queue, and continue the process for each URL in the queue

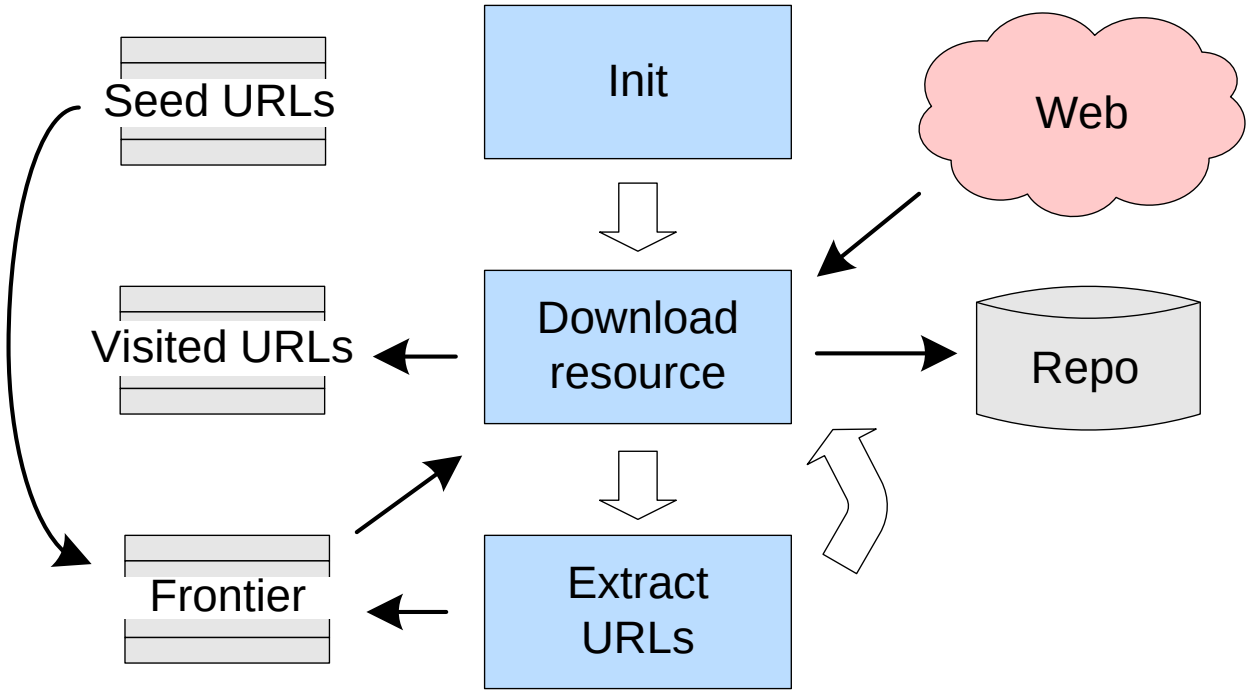


Figure: McCown, *Lazy Preservation: Reconstructing Websites from the Web Infrastructure*, Dissertation, 2007

Problems with Web Crawling

- Slow because crawlers limit how frequently they make requests to the same server (politeness policy)
- Many pages are disconnected from the SCC, password-protected, or protected by robots.txt
- There are an **infinite number** of pages (e.g., calendar) so crawlers limit how deeply they crawl
- Web pages are continually being added and removed
- **Deep web**: Many pages are only accessible behind a web form (e.g., US patent database). Deep web is magnitudes larger than surface web, and 2006 study[1] shows only 1/3 is indexed by big three search engines

[1]He et al., Accessing the deep web, *CACM* 2007

What Counts?

- Many duplicate pages (30% of web pages are duplicates or near-duplicates[1])
  - How do we efficiently compare across a large corpus?
- Some pages change every time they are requested
  - How can we automatically determine what is an insignificant difference?
- Many spammy pages (14% of web pages[2])
  - How can we detect these?

[1]Fetterly et al., On the evolution of clusters of near-duplicate web pages, *J of Web Eng*, 2004
[2]Ntoulas et al., Detecting spam web pages through content analysis, WWW 2006

Some Observations

- Crawling a significant amount of the Web is hard
- Different search engines have different pages indexed, but they don't share these differences with each other (company secret)
- So if we wanted to estimate the Web's size but don't want to try to crawl the Web ourselves, could we use the search engines themselves to estimate the Web's size?

Estimate Web Population

- Lawrence and Giles[1] used capture-recapture method to estimate web page population
  - Submitted 575 queries to sets of 2 search engines
  - S1 = All pages returned by SE1
  - S2 = All pages returned by SE2
  - S1,2 = All pages returned by both SE1 and SE2
  - Size of indexable Web (N) = $S_1 \times S_2/S_{1,2}$
- Estimated size of indexable Web in 1998 = 320 million pages
- Recent measurements using similar methods find lower bound of 21 billion pages[2]

[1]Lawrence & Giles, Searching the World Wide Web, *Science*, 1998
[2]http://www.worldwodewebsize.com/