



# Introduction to Web Science

Intro to Web Science

IB Computer Science (Higher)

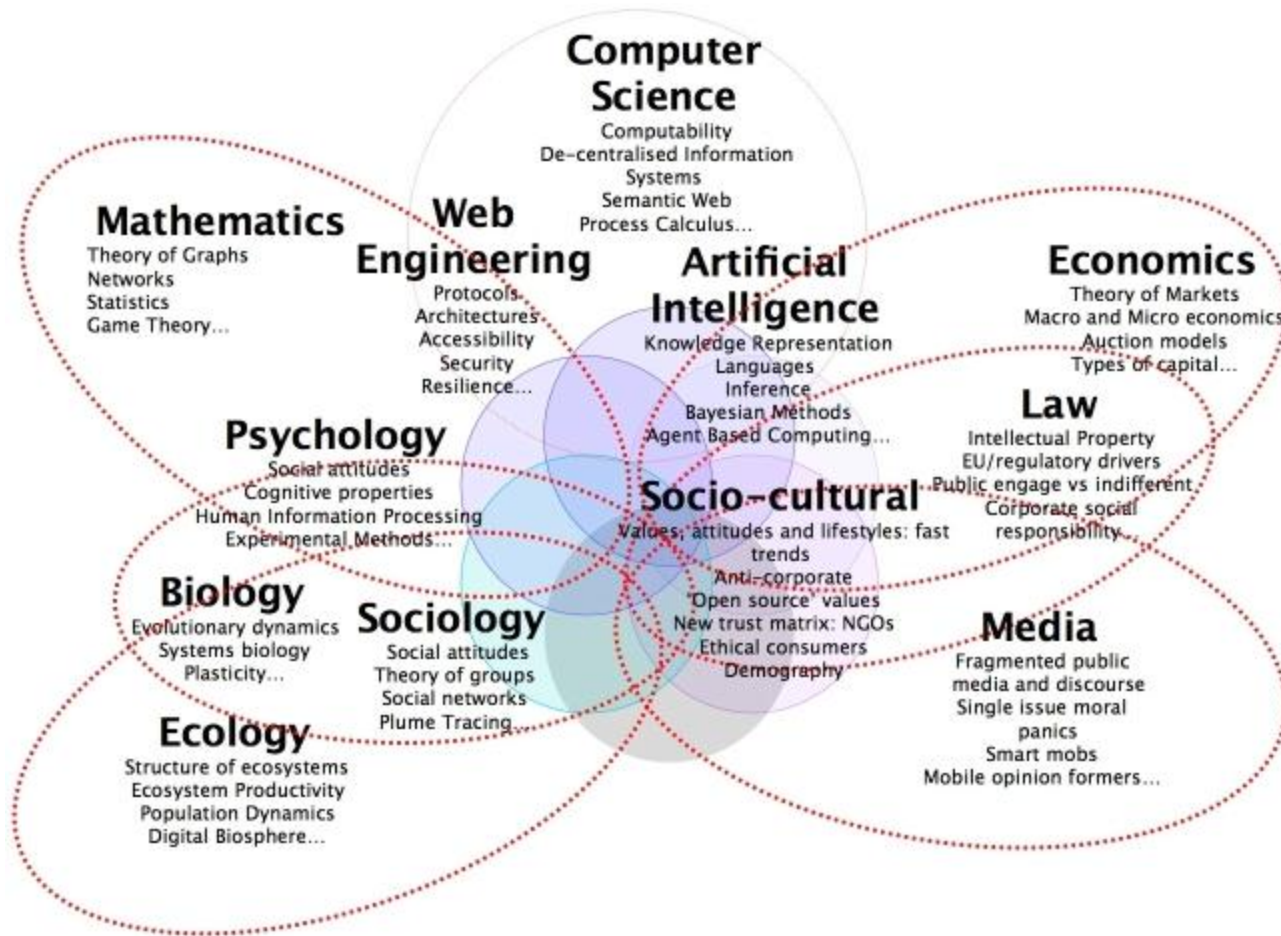
2014

# What is Web Science?

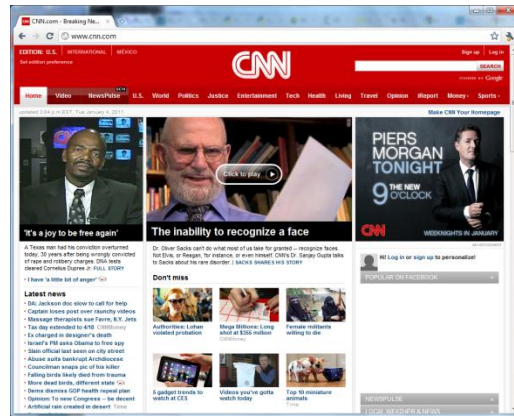
Web Science is the interdisciplinary study of the Web as an entity. It includes studies of the Web's properties, protocols, algorithms, and societal effects.

“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



# How is the Web structured?

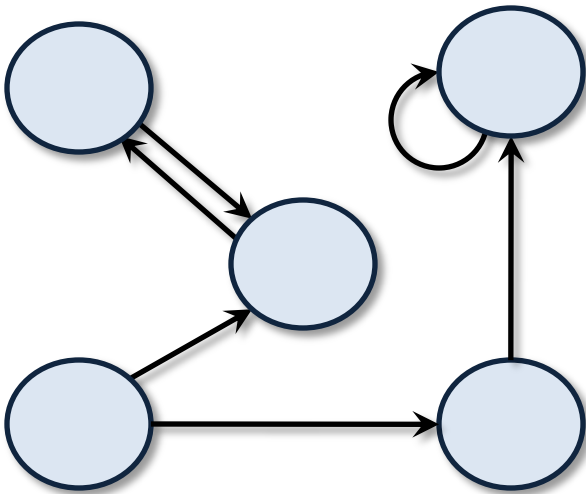


link



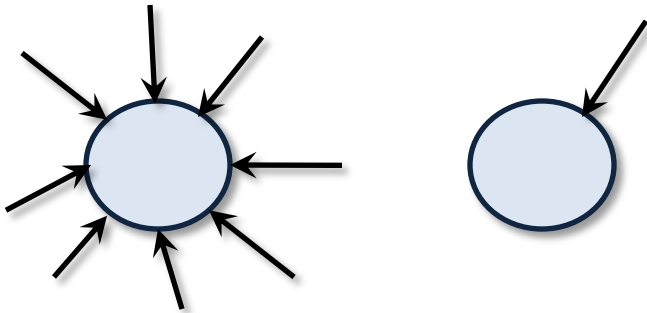
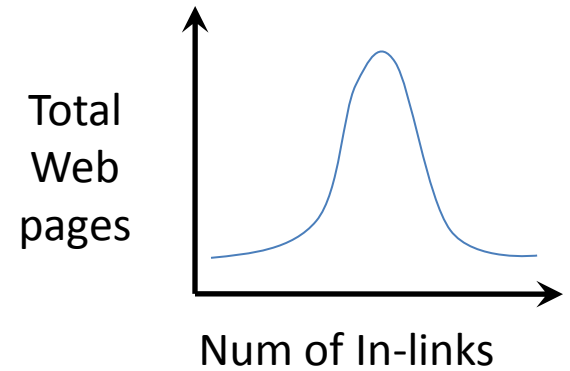
Graph Theory: Pages are nodes & links are directed edges

# Web Graph



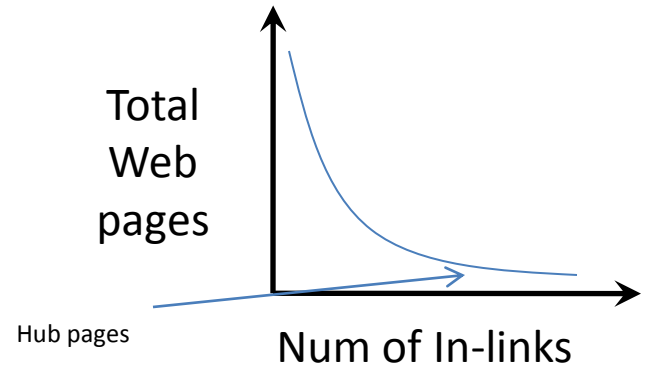
Random Graph

Normal/Gaussian Distribution



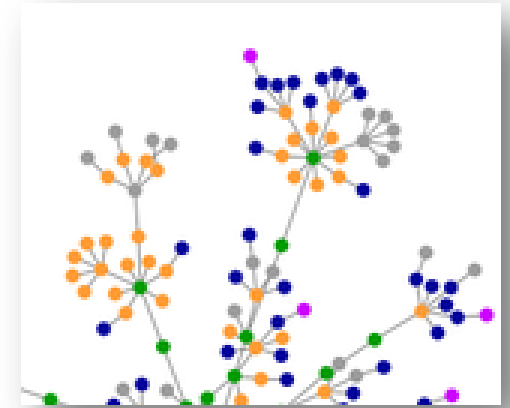
Typical Web Graph

Power-law Distribution

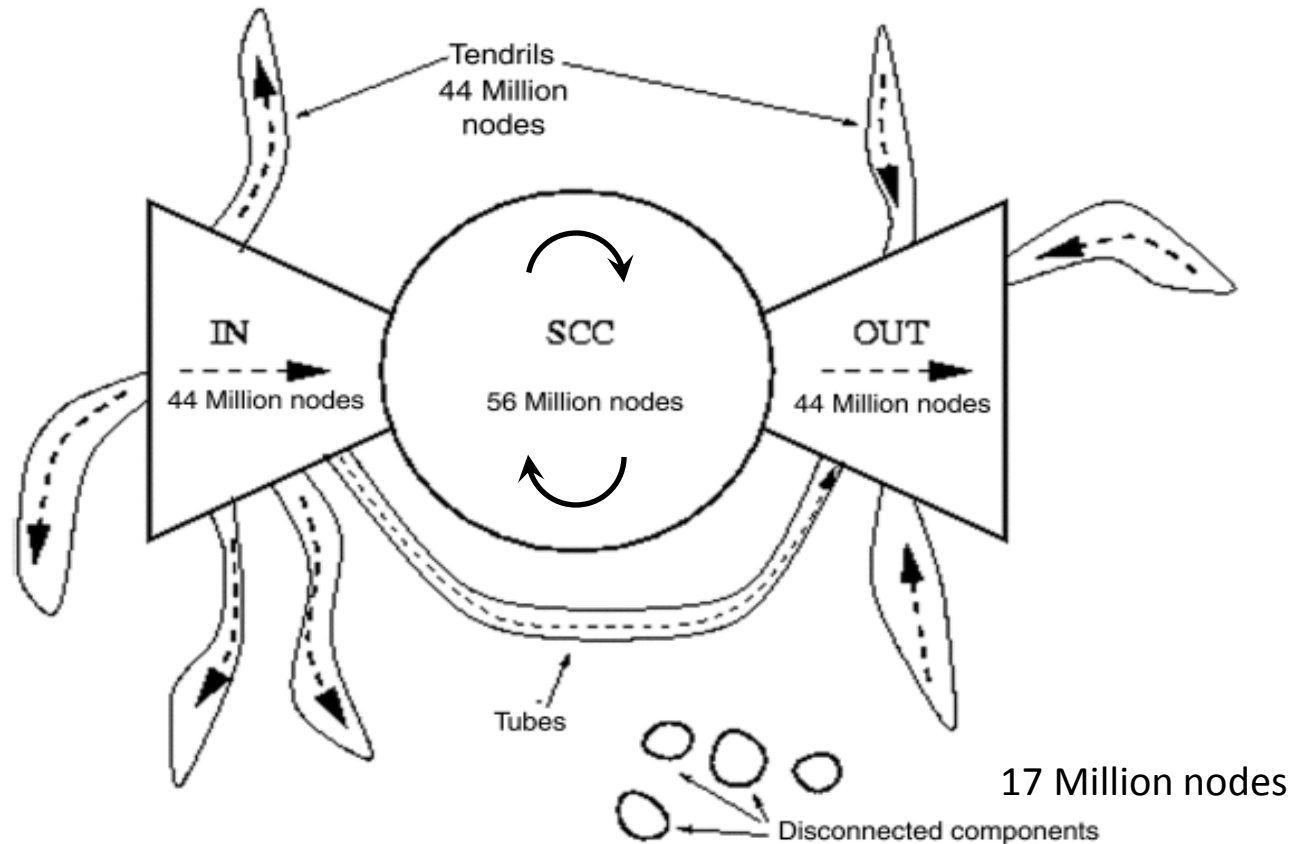


# Small World Network

- 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)

# 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?

The screenshot shows a web browser window displaying a Google Blog post. The browser's address bar shows the URL: `googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`. The page header includes the Google logo and the text "The Official Google Blog | Insights from Googlers into our products, technology, and the Google culture." The main content of the post is titled "We knew the web was big..." and dated "7/25/2008 10:12:00 AM". The text of the post discusses the growth of the web, mentioning that the first Google index in 1998 had 26 million pages, and by 2000 it reached one billion. It states that recently, search engineers were in awe of just how big the web is, with systems processing links hitting a milestone of 1 trillion unique URLs. A blue callout bubble with a white background and a blue border points to the text "1 trillion unique URLs" in the main content area. The right sidebar contains a search box, a "Site Feed" link, a "Make Google your homepage" link, and a "Blog Archive" dropdown menu. The "Labels" section lists various categories with their respective counts: accessibility (28), acquisition (17), ads (85), Africa (2), and Android (1).

Official Google Blog: We ... x

← → ↻ `googleblog.blogspot.com/2008/07/we-knew-web-was-big.html` ☆

Share Report Abuse Next Blog» Create Blog Sign In

The Official **Google** Blog | Insights from Googlers into our products, technology, and the Google culture.

## We knew the web was big...

7/25/2008 10:12:00 AM

We've known it for a long time: the web is big. The first Google index in 1998 already had 26 million pages, and by 2000 the Google index reached the one billion mark. Over the last eight years, we've seen a lot of big numbers about how much content is really out there. Recently, even our search engineers stopped in awe about just **how** big the web is these days -- when our systems that process links on the web to find new content hit a milestone: 1 trillion (as in 1,000,000,000,000) unique URLs on the web at once!

How do we find all those pages? We start at a set of well-connected pages and follow each of their links to new pages. Then we follow the links on those pages to even more pages and so on, until we have more than 1 trillion unique URLs. Even after removing those exact duplicates, we still have a number of individual web pages out there.

1 trillion unique URLs

Search This Blog

Search

powered by Google™

Site Feed

Google

666K readers

BY FEEDBURNER

Make Google your homepage

Blog Archive

Blog Archive ▾

Labels

- [accessibility](#) (28)
- [acquisition](#) (17)
- [ads](#) (85)
- [Africa](#) (2)
- [Android](#) (1)

# 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

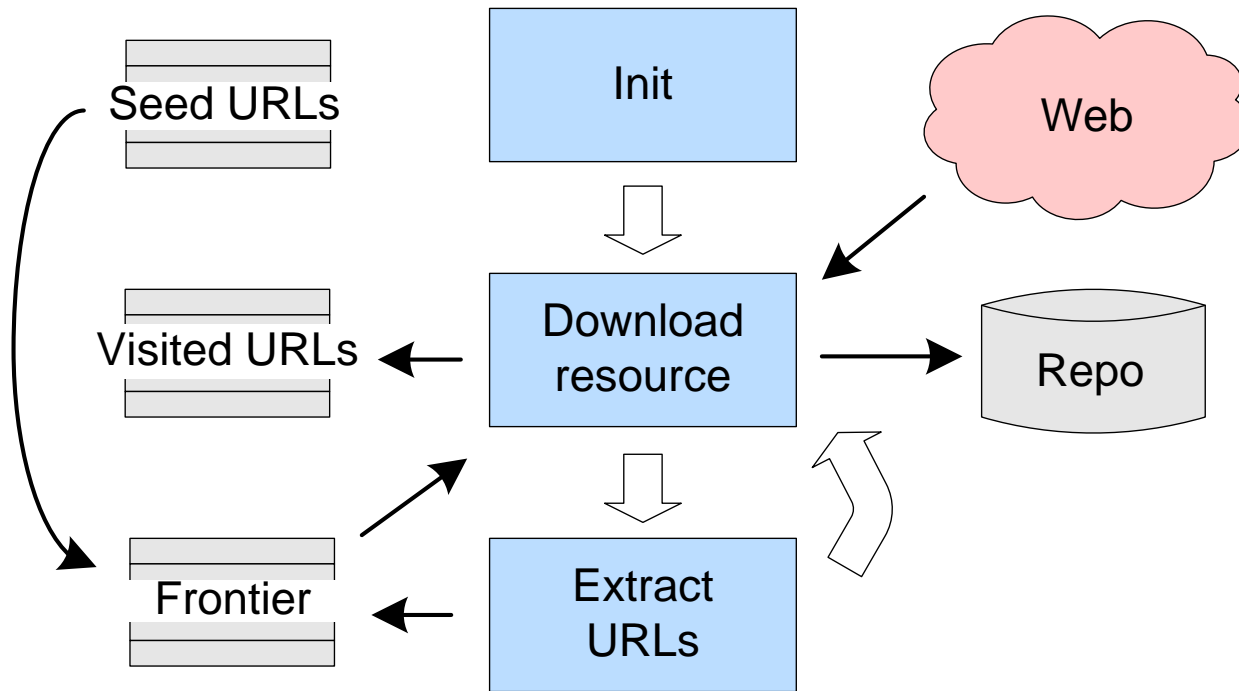


Figure: McCown, *Lazy Preservation: Reconstructing Websites from the Web Infrastructure*, Dissertation, 2007

# 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<sup>1</sup> shows only 1/3 is indexed by big three search engines

<sup>1</sup>He et al., Accessing the deep web, *CACM* 2007

# What Counts?

- Many duplicate pages (30% of web pages are duplicates or near-duplicates<sup>1</sup>)
  - 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<sup>2</sup>)
  - How can we detect these?

<sup>1</sup>Fetterly et al., On the evolution of clusters of near-duplicate web pages, *J of Web Eng*, 2004

<sup>2</sup>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<sup>1</sup> used capture-recapture method to estimate web page population
  - Submitted 575 queries to sets of 2 search engines
  - $S_1$  = All pages returned by SE1
  - $S_2$  = All pages returned by SE2
  - $S_{1,2}$  = 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<sup>2</sup>

<sup>1</sup>Lawrence & Giles, Searching the World Wide Web, *Science*, 1998

<sup>2</sup><http://www.worldwodewebsize.com/>