The World Wide Web

“On the Internet, nobody knows you’re a dog.”

LECTURE 7 – COMPSCI111/111G S1 2017
Today’s lecture

Recap material on the Internet and World Wide Web (WWW)
Understand how the WWW works
Understand how search engines work
The implications of search engines
Recap

In Lecture 4, we saw:

◦ WWW refers to the applications (eg. web pages, email, Skype, Youtube etc) that run on the Internet, which refers to the underlying hardware

◦ The Internet includes the hardware and protocols that transport data from sender to receiver

We’ve already looked at a few WWW applications (eg. email, blogs, instant messaging)
Hypertext

Hypertext is basically text with links
  ◦ Allows associations to be made between pieces of text

Vannevar Bush – “As We May Think” (1945)
  ◦ Bush described a device called a memex, which could store text and links within the text

Ted Nelson – the Xanadu Project (1960s)
  ◦ First computer-based hypertext implementation
  ◦ Although developed in the 1960s, the first public release was in 1998
Multimedia and hypermedia

Multimedia: the integration of many forms of media (text, video, sound, images etc)

Hypermedia: the creation of links between multimedia content
The WWW project

Tim Berners-Lee worked at CERN in the 1980s

Physicists performing research at CERN found it difficult to share their research with each other

Berners-Lee thought he could solve this problem using hypertext and wrote “Information Management: A Proposal” outlining his idea in 1989

- He envisioned a linked information system where pages could be added and accessed by CERN employees
- Pages would be stored on a server
The WWW project

After development in CERN, the first public web server was set up in 1991.

In June 1993, Mosaic was released; the first widely used web browser.

By Oct 1993, there were 500 web servers around the world.
  - By this point, Berners-Lee realised the WWW had to be freely available so he convinced CERN to make the source code public.
The WWW project

In 1994, Berners-Lee established the World Wide Web Consortium (W3C), which creates standards for the WWW
Evolution of the Web

1994: Netscape Communications and Yahoo! founded

1995: first version of Microsoft Internet Explorer released

1998: Google founded

1997-2001: “Dot-com” boom and bust

2004: shift to ‘Web 2.0’ (eg. wikis)
Some terms

**Webpage**: a hypermedia document on the WWW that is usually accessed through a web browser

**Website**: a collection of webpages usually on the same topic or theme

**Web browser**: application software used to access content on the WWW

**Web server**: a computer with software that makes files available on the WWW
Uniform Resource Locator (URL)


Protocol: https
  ◦ Other common protocols: ftp, http

Domain: www.cs.auckland.ac.nz
  ◦ Can be a domain name or an IP address

Path on server: /~andrew/

Resource: teaching.html
HTTP

HyperText Transfer Protocol; used by web browsers to request resources (eg. webpages, images, sounds) from a web server

There’s also HTTPS = HyperText Transfer Protocol Secure

- Encrypts the HTTP connection using TLS (Transport Layer Security)
- Becoming essential for websites to use HTTPS to keep user information secure
Find IP address of www.google.com

CLIENT

GET /index.html HTTP/1.1

HTTP/1.1 200 OK

GET /img/logo.jpg HTTP/1.1

HTTP/1.1 404 NOT FOUND

SERVER
Logging browsing history

A number of computers keep a record of the webpages accessed by a client:
- Web browser
- Computer’s operating system
- ISPs
  - They hold varying amounts of information
  - In Australia, ISPs must retain information about their customers’ web usage for at least 2 years
- The web server
Other parts of the WWW

**Proxy**: sits between client and server so it can intercept and process requests

**Cache**: stores recently requested resources so they can be accessed quickly
- A proxy can use a cache to store recent requests, enabling it to process requests faster

**Firewall**: prevents unauthorised access to a private network
Problems with webpages

Broken links
- Usually the result of a webpage being moved or deleted

No inherent security/tracking/accounting system
- Difficult to have layers of security and a consistent level of security
- Websites rely heavily on ad revenues

No inherent way of indexing information
- Difficult to find information on the web, although search engines help
- Dynamically generated webpages and different file formats (e.g. PDF, archives) also make indexing difficult
Search engines

A website that helps a user to search for information on the WWW

Software indexes content on the web. This index is used to build a list of results based on the search terms entered by the users

- **Indexing**: organising data so that it is easier to search

Popular search engines include:

- Google
- Bing
- Yahoo search
- DuckDuckGo
Search engines
How do search engines work?

Spiders crawl across the WWW to scan webpages
- Spiders are programs that follow links and gather information from webpages

The search engine’s index is updated with information gathered by the spiders
How do search engines work?

User enters a search term

The search engine uses algorithms to find the most relevant results in its index

◦ These algorithms are secret and highly complex
◦ They use a number of criteria, such as keywords and popularity, to determine a page’s relevance to the user

Search engine gives the user a list of results

◦ This list is compiled from billions of webpages in a couple of seconds!
Can we trust search engines?

Bias in the results?
- Since search algorithms are secret, we have to trust that they operating fairly
- Effect of filtering on search results (eg. DMCA, images of child abuse)

Advertising plays a big role in how search engines operate
- Search engines make money from advertising
- Companies misuse search engines to get a competitive edge: NakedBus using ‘inter city’ on Google Adwords (a good summary can be found here)
Can we trust search engines?

The right to be forgotten (R2BF)

- In 2014, European Court of Justice decided R2BF meant Google has to remove out-of-date search results when requested by individuals
- A good summary can be found here
- In Europe, the General Data Protection Regulation 2016 contains a more limited ‘right to erasure’

R2BF helps an individual to preserve their privacy

However, the R2BF distorts search results and could be abused (e.g., a businessman wanting news articles removed from search results)
Filter bubble

Occurs when a search algorithm offers personalised results, which limits the diversity of information presented to the user

- Examples include Facebook’s News Feed and Google’s personalised search results

Personalised search results can help people to find relevant information

However, it also risks isolating people within their own bubble of information
Privacy

Search engines are gathering vast amounts of information about our searches and ourselves

- This information is generally used for advertising purposes

Can we trust private companies to treat our information with care? To keep it secure? To not sell it to others without consent?

While you can search anonymously, search history can be used to identify individuals

- A reporter used a person’s anonymised search history to track them down – article [here](#)
Summary

The WWW was designed to be a system to share information
  ◦ It has become a system for creating and sharing a variety of content
  ◦ Key protocol on the WWW is HTTP

Search engines use an index of the WWW to provide results based on search terms

Issues around search engines
  ◦ Bias
  ◦ Protecting privacy (eg. R2BF)
  ◦ Use of personal information for advertising
  ◦ Filter bubbles