Lecture 17: (World Wide) Web

- a way to connect computers that provide information (servers) with computers that ask for it (clients like you and me)
 - uses the Internet, but it's not the same as the Internet
- URL (uniform resource locator, e.g., http://www.amazon.com)
 - a way to specify what information to find, and where
- HTTP (hypertext transfer protocol)
 - a way to request specific information from a server and get it back
- HTML (hyptertext markup language)
 - a language for describing information for display
- browser (Firefox, Safari, Chrome, Edge, ...)
 - a program for making requests, and displaying results
- embellishments
 - pictures, sounds, movies, ...
 - loadable software
- the set of everything this provides

Web history

- 1989: Tim Berners-Lee at CERN
 - a way to make physics literature and research results accessible on the Internet
- 1991: first software distributions
- Feb 1993: Mosaic browser
 - Marc Andreessen at NCSA (Univ of Illinois)



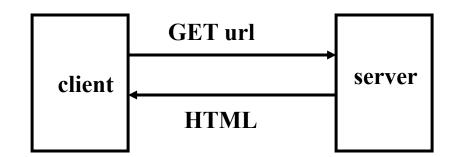
- first commercial browser
- technical evolution managed by World Wide Web Consortium
 - non-profit organization at MIT, Berners-Lee is director
 - official definition of HTML and other web specifications
 - see www.w3.org



HTTP: Hypertext transfer protocol

- What happens when you click on a URL?
- client opens TCP/IP connection to host, sends request

- server returns
 - header info
 - HTML



- server returns text, which can be dynamically created as needed
 - can contain encoded material for images, music, video (MIME format)
- URL format

```
service://hostname/filename?other_stuff
```

- filename?other stuff part can encode
 - data values from client (forms)
 - request to run a program on server (cgi-bin)
 - anything else

Embellishments

- original design of HTTP just returns text to be displayed
- MIME format for pictures, sound, video, ...
 - helpers or plug-ins display non-text content
- forms filled in by user
 - needs a program on the server to interpret the information (cgi-bin)
- cookies to remember information on client
 - HTTP is stateless: server doesn't save anything from one request to next
 - cookies are a way to remember information at the client
- Javascript: download code to run on the client

Forms and CGI programs

- "common gateway interface"
 - standard way to request the server to run a program
 - using information provided by the client via a form
- if the target file on the server is an executable program
- and if it has the right properties and permissions
 - e.g., in /cgi-bin directory and executable
- then run it on the server to produce HTML to send back to the client
 - using the contents of the form as input
 - output depends on client request: created on the fly, not just a file
- CGI programs can be written in any programming language
 - e.g., Python, Java, ...

Cookies

- HTTP is <u>stateless</u>: it doesn't remember from one request to the next
- cookies are intended to deal with stateless nature of HTTP
 - remember preferences, manage "shopping cart", etc.
- cookie: one chunk of text sent by server to be stored on the client
 - stored in browser while it is running (transient)
 - stored in client file system when browser terminates (persistent)
- when the client reconnects to same domain,
 browser sends the cookie back to the server
 - sent back verbatim; nothing added
 - sent back only to the same domain that sent it originally
 - contains no information that didn't originate with the server
- in principle, pretty benign
- but pervasively used to monitor browsing, for commercial purposes

Cookie crumbs

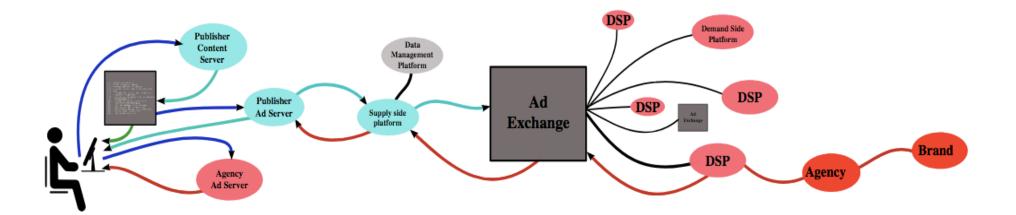
- fetch a page from xyz.com
 - it contains
 - this causes a page to be fetched from DoubleClick.com (part of Google)
 - which now knows your IP address and what page you were looking at
- DoubleClick sends back (or arranges for) a suitable advertisement
 - with a cookie that identifies "you" at DoubleClick
- next time you fetch <u>any</u> page that contains a DoubleClick.com image
 - the most recent DoubleClick cookie is sent back to DoubleClick
 - DoubleClick now knows even more about you
 - the set of sites and images that you are viewing is used to
 - · update the record of where you have been and what you have looked at
 - · send further targeted advertising
 - · allow more cookies from advertisers

Advertising marketplace

- When you use a browser to view a web page...
- the web page "publisher" notifies advertising exchanges that advertising space on that page is available
 - publishers are often social media or entertainment or news sites
 - exchanges include Google Ad Manager, Microsoft Xandr, Yahoo, ...
- the publisher provides information about you to the ad exchange
 - past online activity, viewing and shopping habits, geographic location, demographics
 - (probably) not your actual identity
- advertising exchanges tell potential advertisers about you
- advertisers bid on the ad space
 - amount depends on your attributes and location, advertiser's budget, etc.
- winners' advertisements are inserted into the page that you see
- elapsed time: 10-100 milliseconds

Who's involved?

- publisher: integrates advertisements into its online content
- advertiser: provides the advertisements to be displayed on the publisher's content
- advertising agencies: generate and place the ad copy
- data supplier: provide information about the viewer
- ad server: delivers the ad and tracks statistics



Cookies are not the only tracking mechanism

- 3th party cookies are decreasing in value as more browsers block them by default
- Alternatives:
- JavaScript
 - potentially continuous monitoring and reporting of activity on a page
- web bugs, web beacons, single-pixel gifs
 - tiny images that report the use of a particular page
 - these can be used in mail messages, not just browsers
- HTML canvas fingerprinting
 - uses subtle differences in browser behavior to distinguish users

Javascript

- programming language loosely in the C family (surface syntax similar)
 - (no relationship to Java)
 - compiled into instructions for a virtual machine like the Toy machine on steroids
 - instructions are interpreted by a virtual machine in browser
- most common use is embedded in web pages, running in browser
 - can also run standalone

- can interact with browser to see what is displayed, change what is displayed
 - can watch events like clicks, mouse motion, ...
 - can send and receive data from network (with restrictions)
 - can load more Javascript from network (with restrictions)

What does Facebook know about you?

- https://www.cnbc.com/2018/03/27/facebook-knows-a-lot-about-me.html
- It can recognize my face
- It knows every ad topic I've ever clicked
- It has a list of every company that has my contact information from the ads I've clicked
- It has a list of every contact in my phone book
- It knows every social event I was invited to and/or attended through Facebook
- It has a log of every friend I have on Facebook and when we became friends
- It knows every time I logged in
- It has a copy of my timeline going back to the time I joined
- It knows my major life events
- It knows every video I've watched on Facebook
- It knows exactly where I was
- It has old messages
- It has a copy of every photo I've ever uploaded