

# The World Wide Web

- **what it is**
- **a brief history**
- **how it works**
- **how advertising works**
- **how other things work**
- **technical issues**
- **political / legal / social / economic / jurisdictional issues**

# (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 (hypertext markup language)**
  - a language for describing information for display
- **browser (Firefox, Safari, Internet Explorer, Opera, Chrome, ...)**
  - a program for making requests, and displaying results
- **embellishments**
  - pictures, sounds, movies, ...
  - loadable software
- **the set of everything this provides**

# HTTP: Hypertext transfer protocol

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

```
GET /filename HTTP/1.0
```

- server returns
  - header info
  - HTML



- since server returns the text, it can be created as needed
  - can contain encoded material of many different types (MIME)

- 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**
- **now includes pictures, sound, video, ...**
  - need helpers or plug-ins to display non-text content  
e.g., GIF, JPEG graphics; sound; movies
- **forms filled in by user**
  - need 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
- **active content: download code to run on the client**
  - Javascript
  - Java applets
  - plug-ins
  - ActiveX

# 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 server is an executable program**
- **and it has the right properties and permissions**
  - e.g., in /cgi-bin directory and executable
- **then run it on server to produce HTML to send back to 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**
  - Perl, Python, PHP, Java, Ruby, ...

# Cookies

- **HTTP is stateless: doesn't remember from one request to next**
- **cookies 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 client**
  - stored in browser while it is running (transient)
  - stored in client file system when browser terminates (persistent)
- **when 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 heavily used to monitor browsing habits, for commercial purposes**

# Advertising

- **advertising exchanges**
  - Yahoo Right Media, Doubleclick Ad Exchange, Facebook Atlas ...
- **a person uses a browser to request a web page**
- **web page "publisher" notifies exchange that advertising space on that page is available**
  - publishers are typically portals or entertainment and news sites
  - publisher provides information about the person: past online activity, viewing and shopping habits, geographic location, demographics  
probably not actual identity (?)
- **advertisers bid on the ad space**
  - amount depends on person's attributes and location, advertiser's budget, etc.
- **winner's advertisement is inserted into the page**
- **elapsed time: 10-100 milliseconds**
- **this happens for multiple advertisements on one page**

# Cookies are not the only tracking mechanism

- **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
- **Flash cookies ("local shared object")**
  - cookie-like mechanism used by Flash
- **"super cookies"**
  - e.g., Verizon's X-UIDH HTTP header on cellphones
- **HTML canvas fingerprinting**
  - uses subtle differences in browser behavior to distinguish users
  
- **defenses:**
  - addons like Adblock, FlashBlock, Cookie Monster, Ghostery, NoScript**



# Plug-ins, add-ons, extensions, etc.

- **programs that extend capabilities of browser (and other programs)**
  - browser provides an API and a protocol for data exchange
  - extension focuses on specific application area
    - e.g., documents, pictures, sound, movies, scripting language, ...
  - may exist standalone as well as in plug-in form
  - e.g., Acrobat Reader, Flash, Quicktime, Windows Media Player, ...
- **scripting languages interpret downloaded programs**
  - Javascript
    - compiled into instructions for a virtual machine
      - (like the Toy machine on steroids)
    - instructions are interpreted by virtual machine in browser

# Potential security & privacy problems

- **attacks against client**

- release of client information

- cookies: client remembers info for subsequent visits to same server

- adware, phishing, spyware, viruses, ...

- spyware: client sends info to server upon connection (Sony, ...)

- often from unwise downloading

- buggy/misconfigured browsers, etc., permit vandalism, theft, hijacking, ...

- **attacks against server**

- client asks server to run a programs when using cgi-bin

- server-side programming has to be careful

- buggy code on server permits break-in, theft, vandalism, hijacking, ...

- denial of service attacks

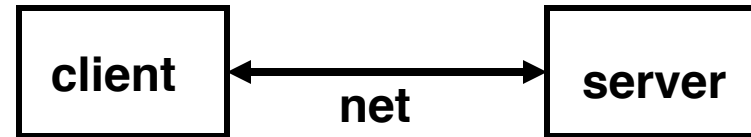
- **attacks against information in transit**

- eavesdropping

- encryption helps

- masquerading

- needs authentication in both directions



# Privacy on the Web

- **what does a browser send with a web request?**
  - IP address, browser type, operating system type
  - referrer (URL of the page you were on)
  - cookies
- **what do "they" know about you?**
  - whatever you tell them, implicitly or explicitly (e.g., Facebook)
  - public records are really public
  - lots of big databases like phone books
  - log files everywhere
  - aggregators collect a lot of information for advertising
  - spyware, key loggers and similar tools collect for nefarious purposes
  - government spying is everywhere
- **who owns your information?**
  - in the USA, they do
  - less so in the EU

# Defenses

- **use strong passwords; don't share them across important accounts**
- **use 2-factor identification when available (e.g., Duo)**
- **cookies off, spam filter on, Javascript limited**
- **turn off previewers and HTML mail readers**
- **anti-virus software on and up to date**
  - turn on macro virus protection in Word, etc.
- **run spyware detectors**
- **use a firewall**
- **try less-often targeted software**
- **be careful and suspicious all the time**
  - don't view attachments from strangers
  - don't view unexpected attachments from friends
  - don't just read/accept/click/install when requested
  - don't install file-sharing programs
  - be wary when downloading software

