How the Internet Works

Steven M. Bellovin
Department of Computer Science, Columbia University
https://www.cs.columbia.edu/~smb

Goals for Lawyers

- Red-titled slides summarize tentative legal conclusions
- Who knows what?
- How do they know it?
- What is the legal environment of that knowledge?
  - Private data
  - Third party doctrine
  - Wiretap Act
- Are there relevant regulatory issues?

What is the Internet Made of?

- Computers
  - Servers
  - Clients
  - Phones
  - “Things”
- Routers—specialized computers that forward “packets”
  - Packets are fragments of messages
- Links—WiFi, Ethernet, fiber, etc. The Internet was designed to run over *anything*

Fibers

- Each cable has many pairs of *strands*
- Each strand carries many *wavelengths* (aka “colors” or “lambdas”)
  - A new trans-Pacific fiber has six pairs of strands
  - Each strand carries 100 wavelengths
  - Each wavelength has a bandwidth of 100G bps
  - Total capacity: 60 terabits/second
- Each wavelength can carry many different circuits
- Each Internet circuit carries packets for many different conversations

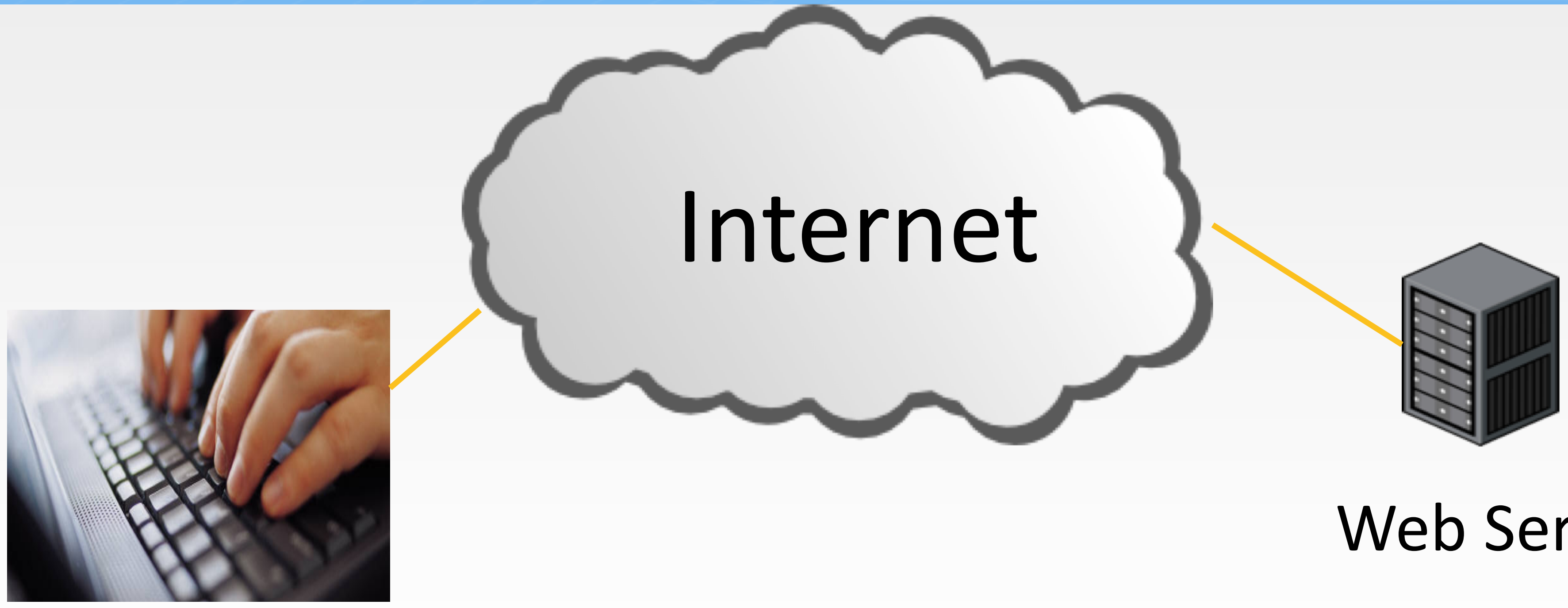
WiFi

- Used in public spaces and private residences
  - Some use in business, but wired Ethernet is more common for desktops
- Range: about 100 meters
- Security: WEP is obsolete and insecure; WPA2 is quite good—and in public, all bets are off.

A Look at Common Applications

- Web browsing
- Email
- The Cloud
- *Caution: all of this is simplified—and arguably oversimplified*

How the Web Appears to Users

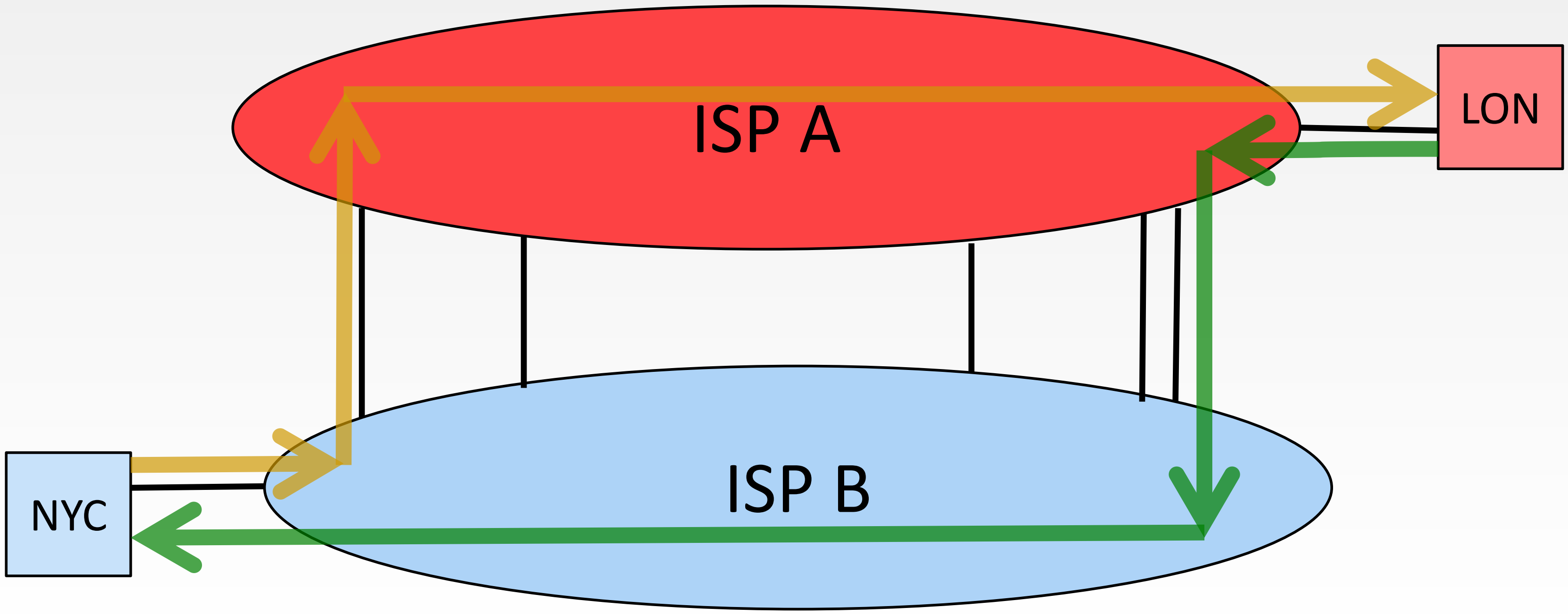


Internet

Web Server

Web Browser

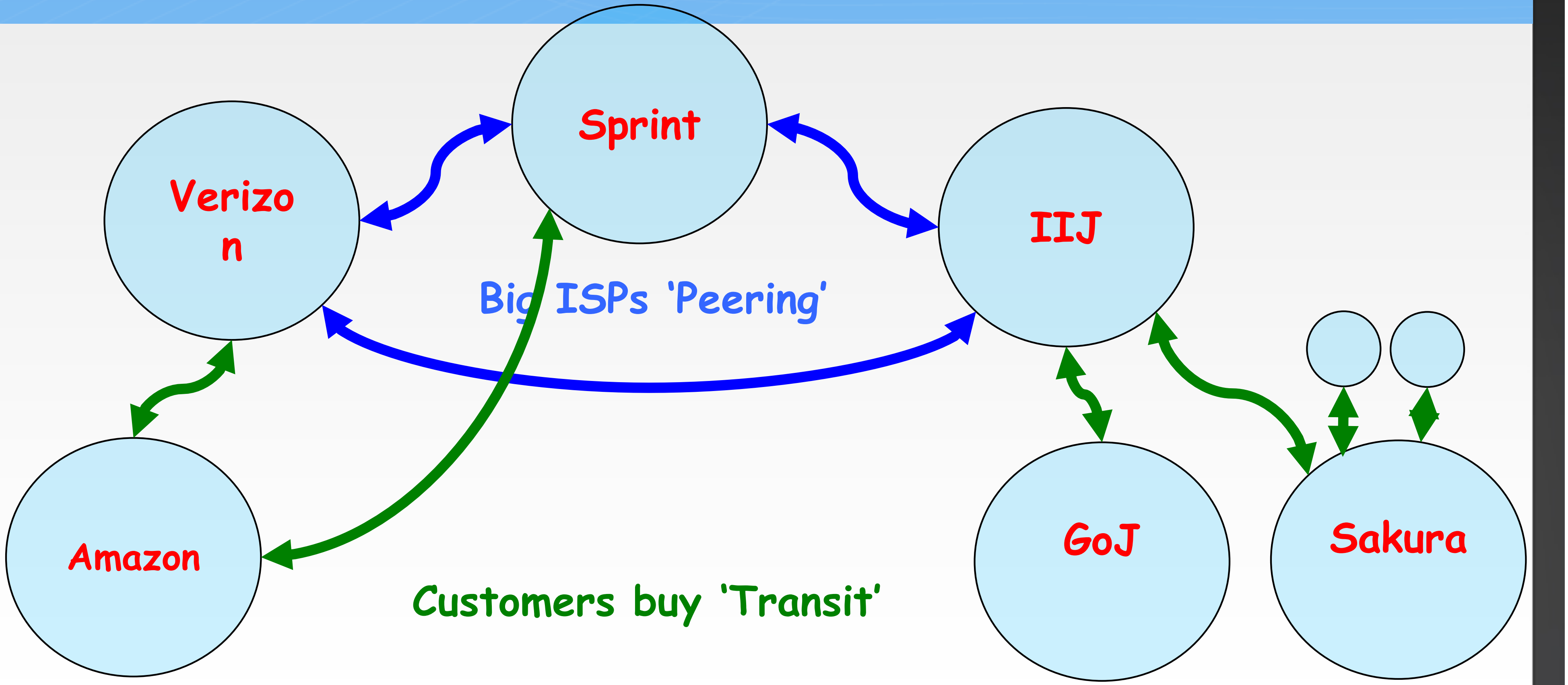
The Internet Has Structure: Multiple ISPs that Interconnect at Multiple Points



Wiretapping the Internet

- Tapping the “backbone” is very difficult—the forward and reverse directions of a conversation generally follow different paths
- Consequence: wiretap orders should be served on edge providers

Routing Between ISPs



Sprint

Verizo
n

IIJ

Big ISPs 'Peering'

Amazon

GoJ

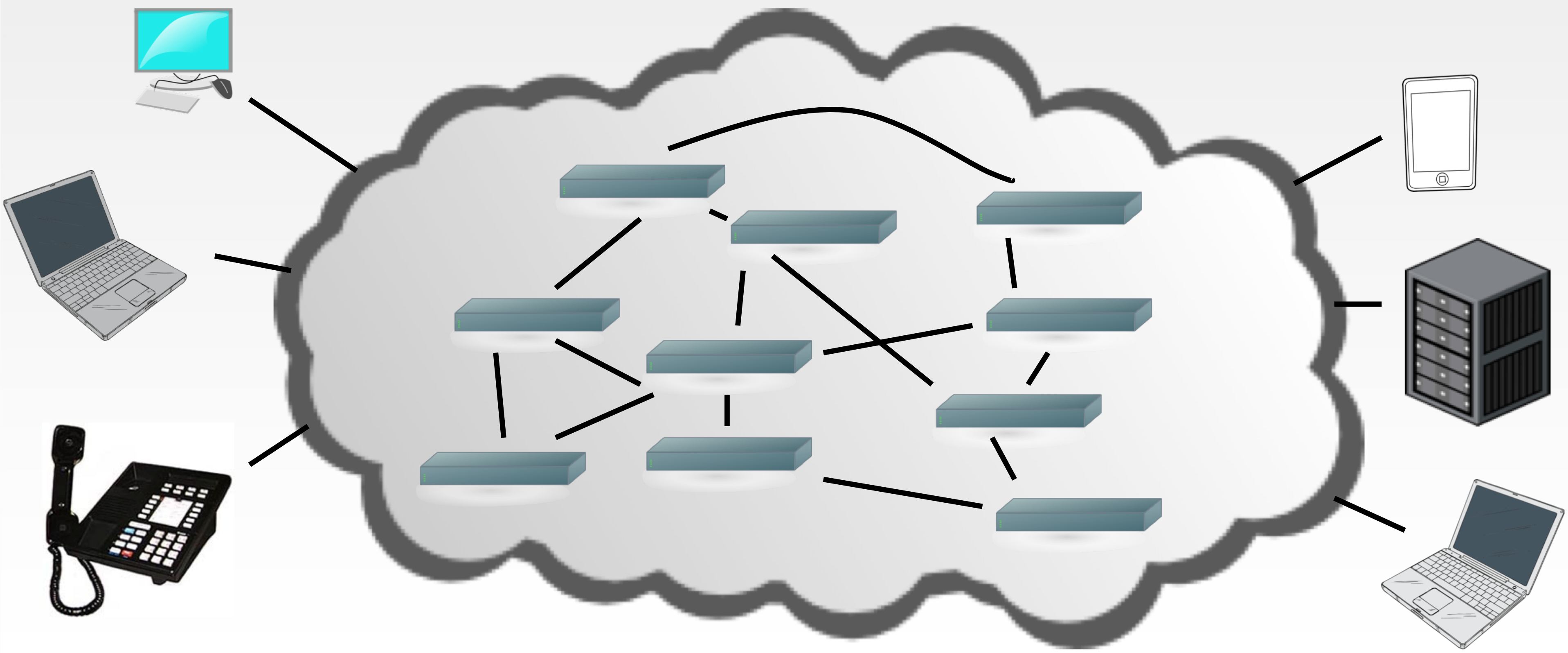
Sakura

Customers buy 'Transit'

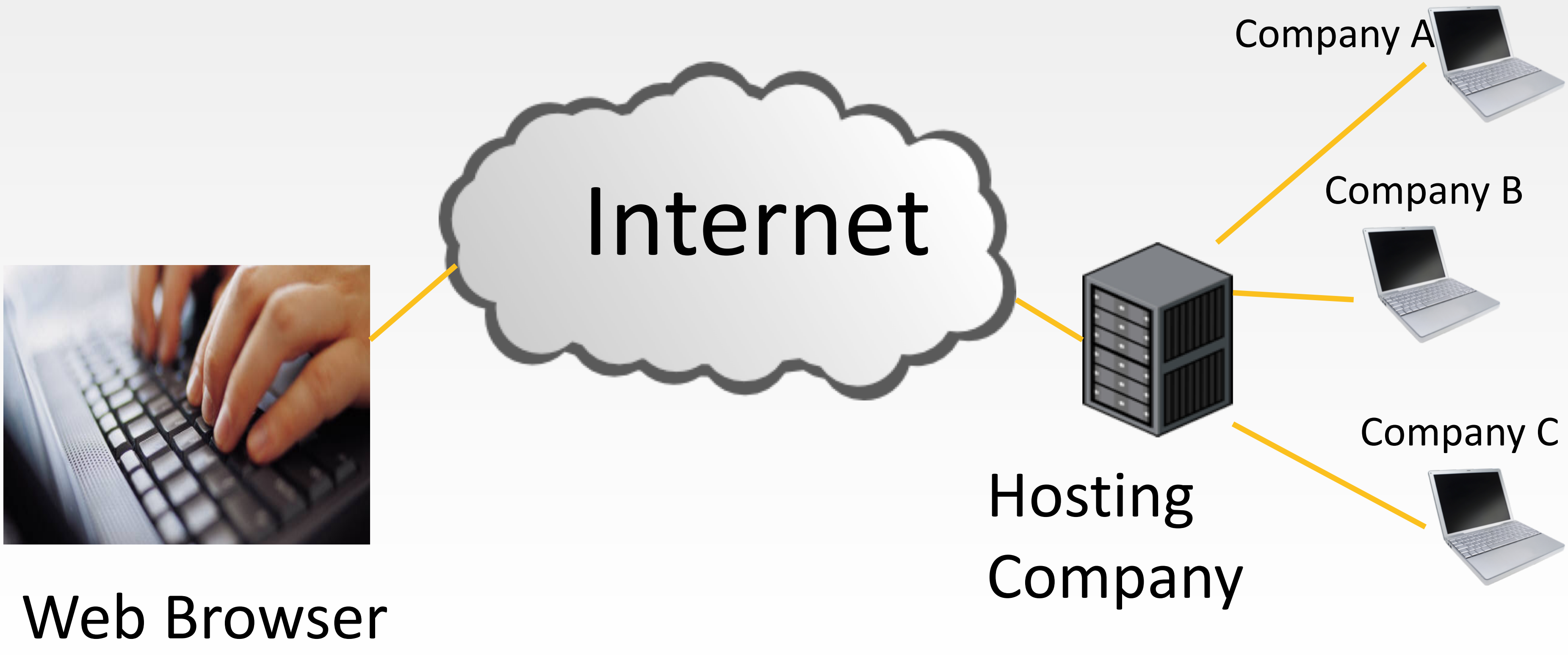
Some Regulatory Issues

- Contracts are generally bilateral and confidential
- It’s hard to know a priori if there are antitrust issues
- Net neutrality was primarily about putting restrictions on these contracts

Each ISP Has Structure: Many Routers



Hosting Services



Internet

Web Browser

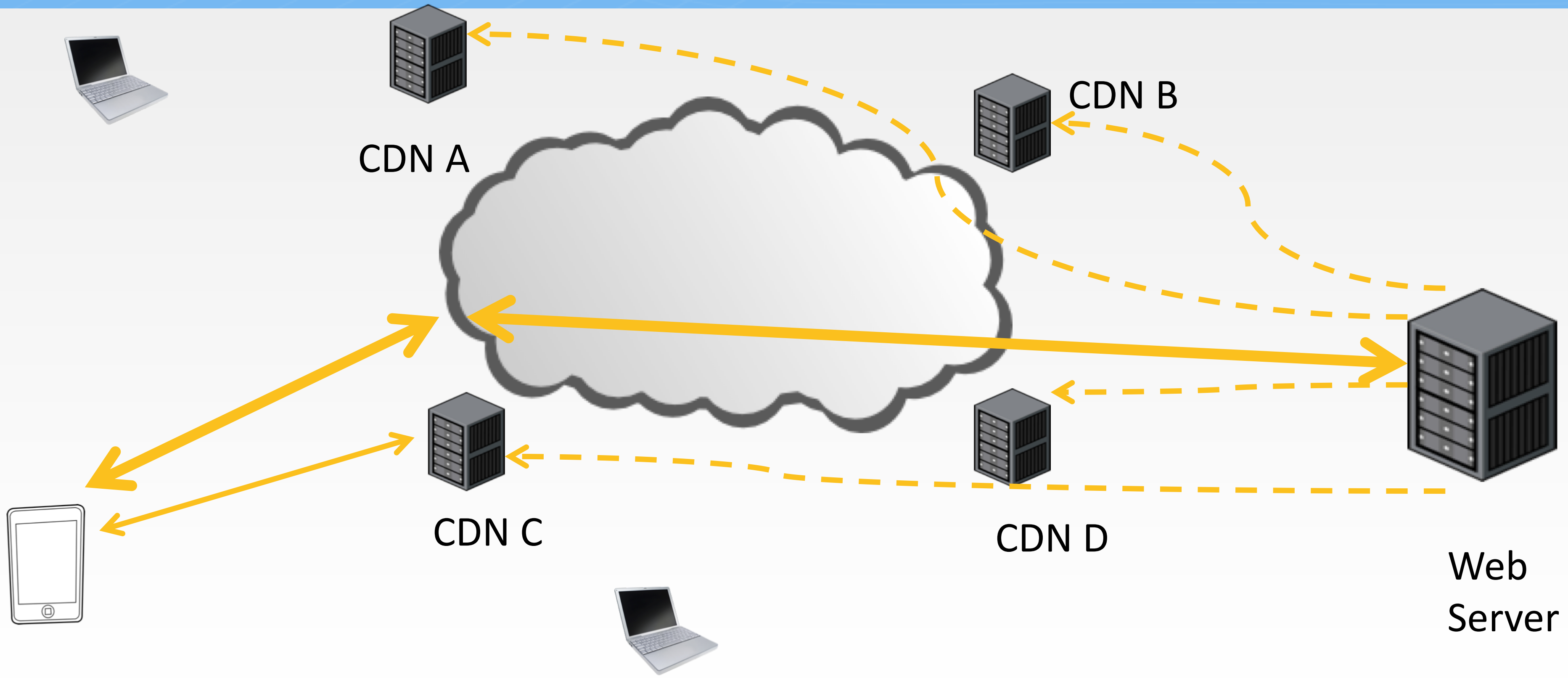
Hosting Company

Company A

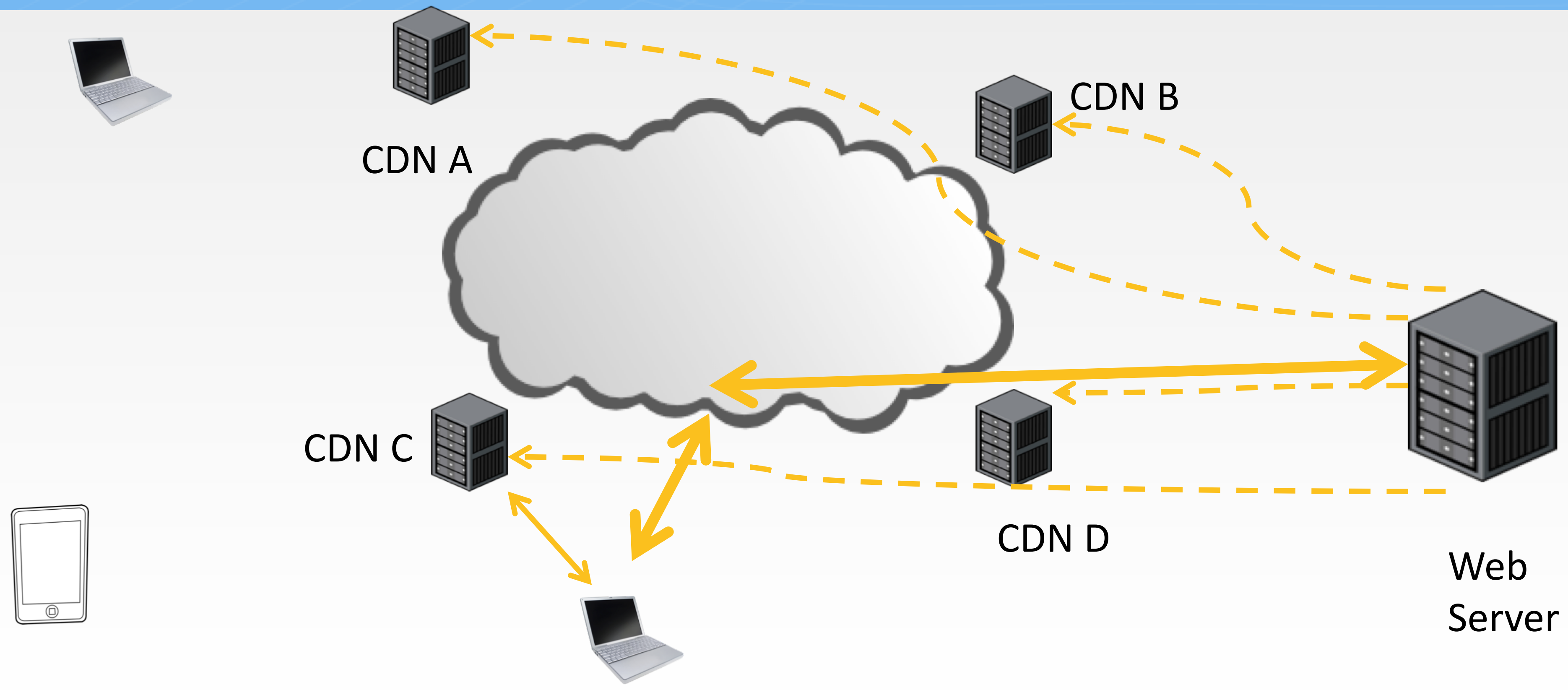
Company B

Company C

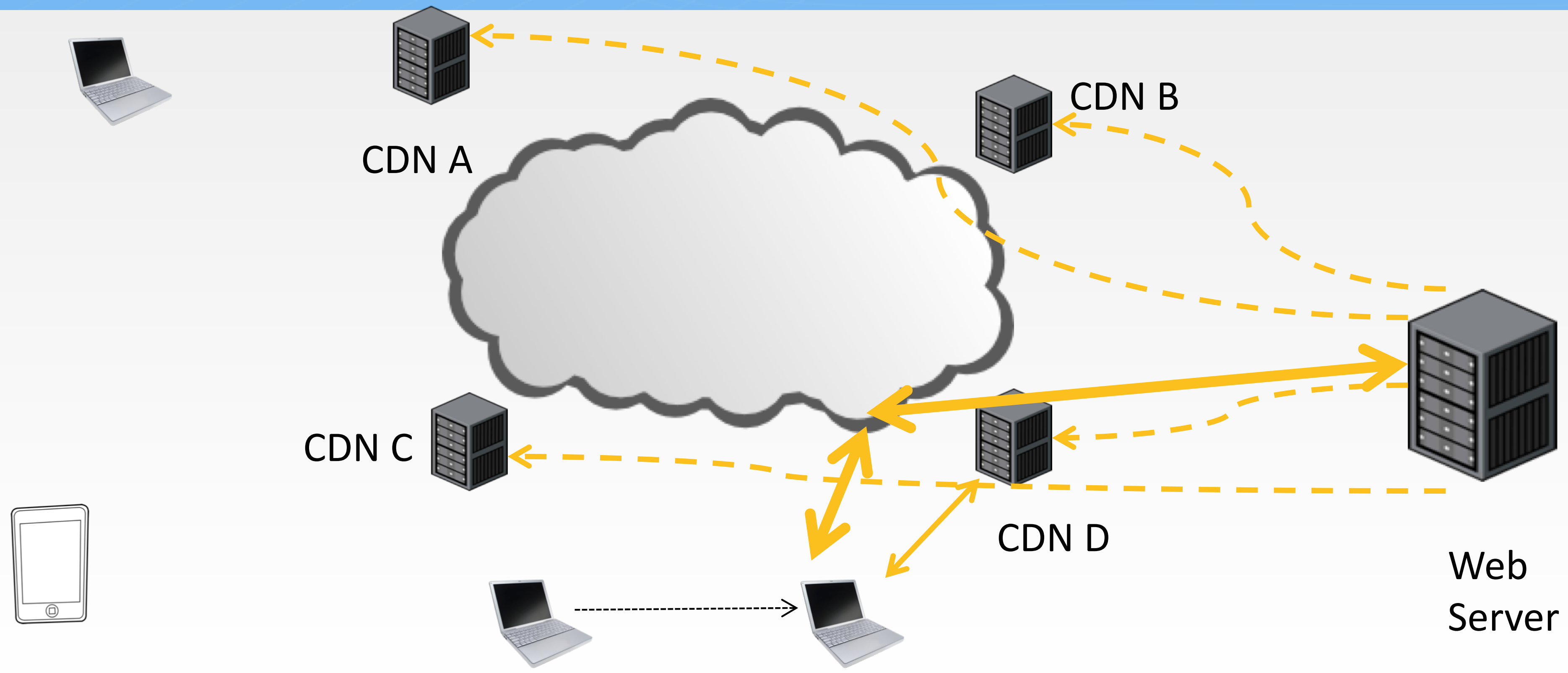
Content Distribution Network



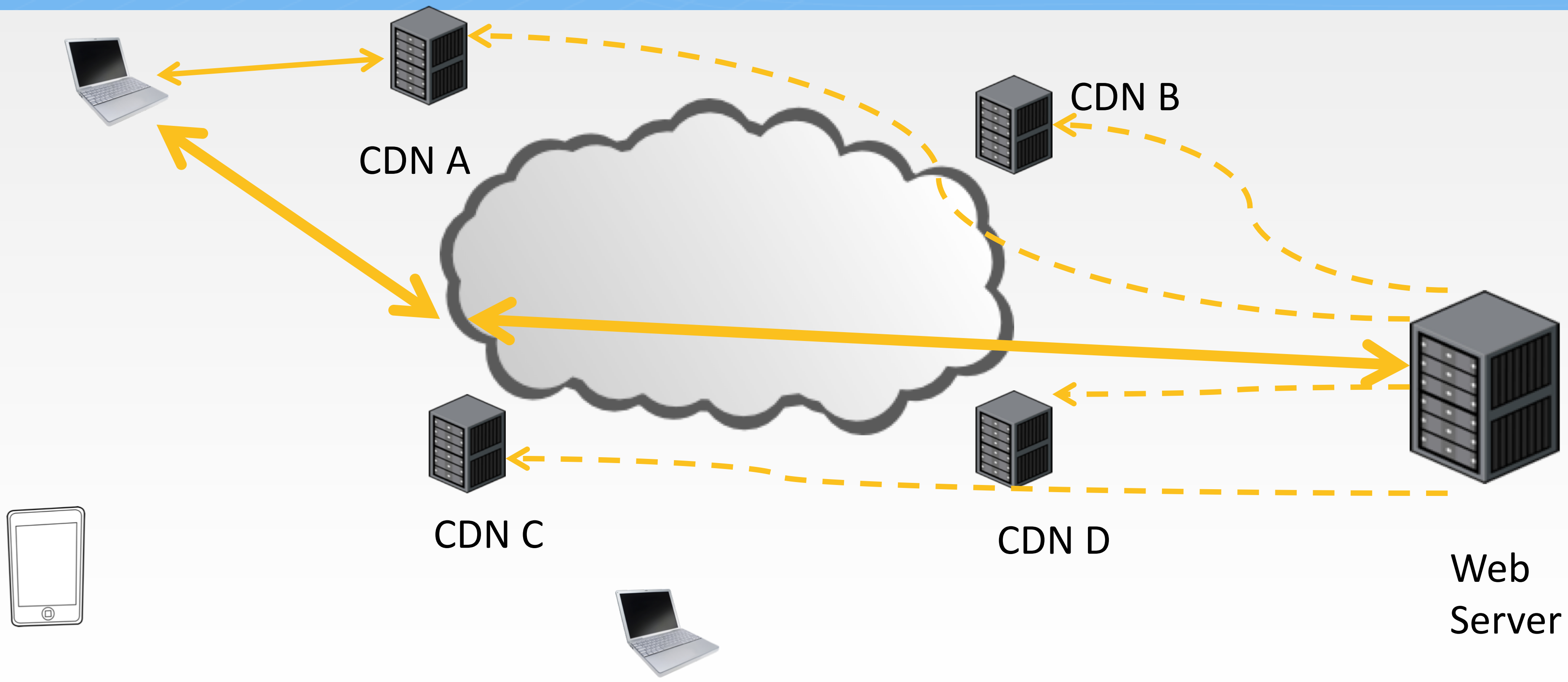
Content Distribution Network



Content Distribution Network



Content Distribution Network



CDN Example: www.supremecourtus.gov

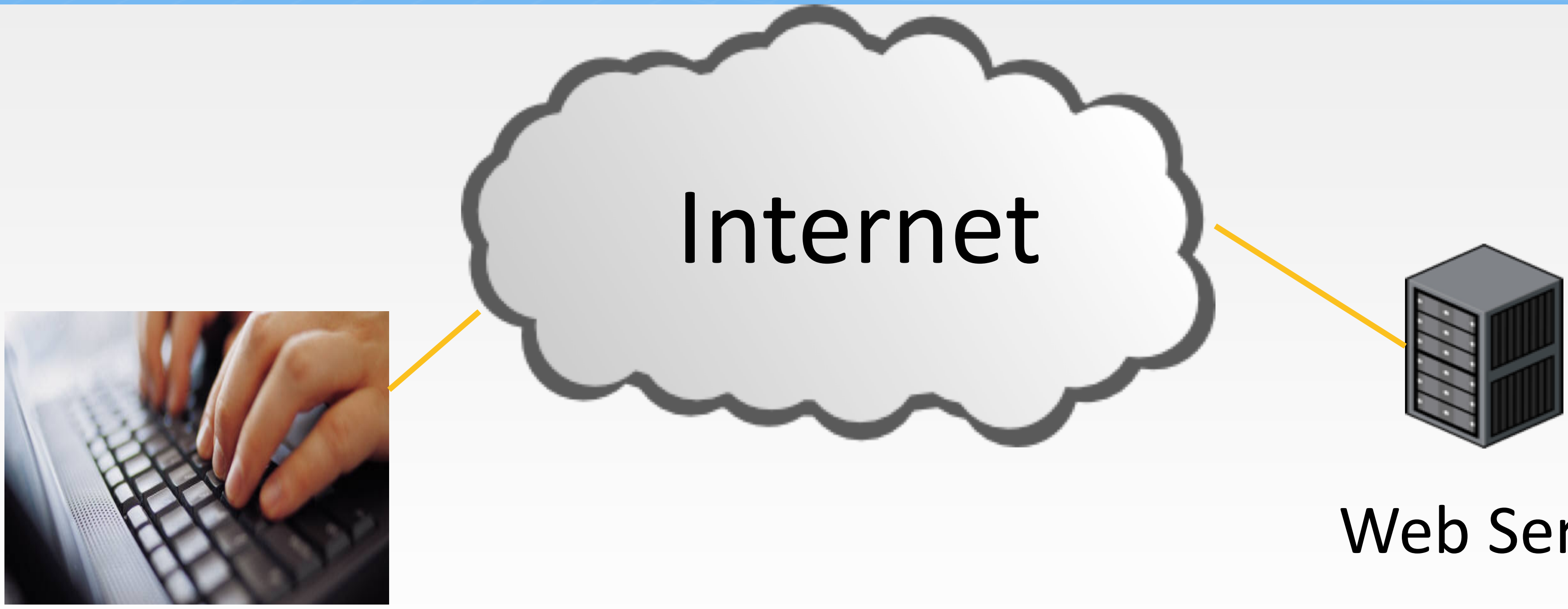
www.supremecourt.gov is an alias for a1042.b.akamai.net; Akamai is a prominent CDN operator

New York	24.143.200.48
Ashburn, Va	23.15.9.144
Atlanta	208.44.23.57
San Francisco	216.156.149.106
Boston	207.86.164.89

Multiple Parties

- There are multiple parties involved in most web transactions
- Consequently, multiple know who is visiting what pages and hence when is retrieving what content
- However, determining who has the information for any given transaction is not always easy

Which is the Browser; Which is the Server?

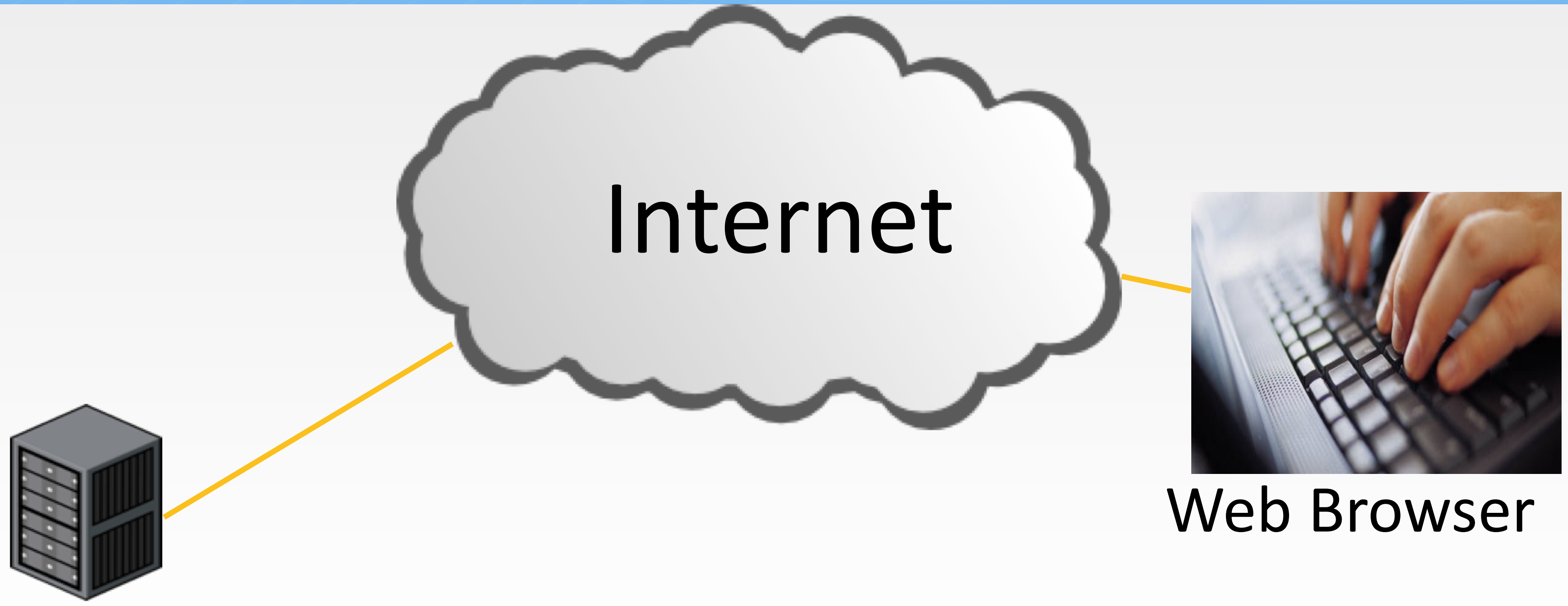


Internet

Web Server

Web Browser

Architecturally, They're the Same—What Matters is the Software They Run



Internet

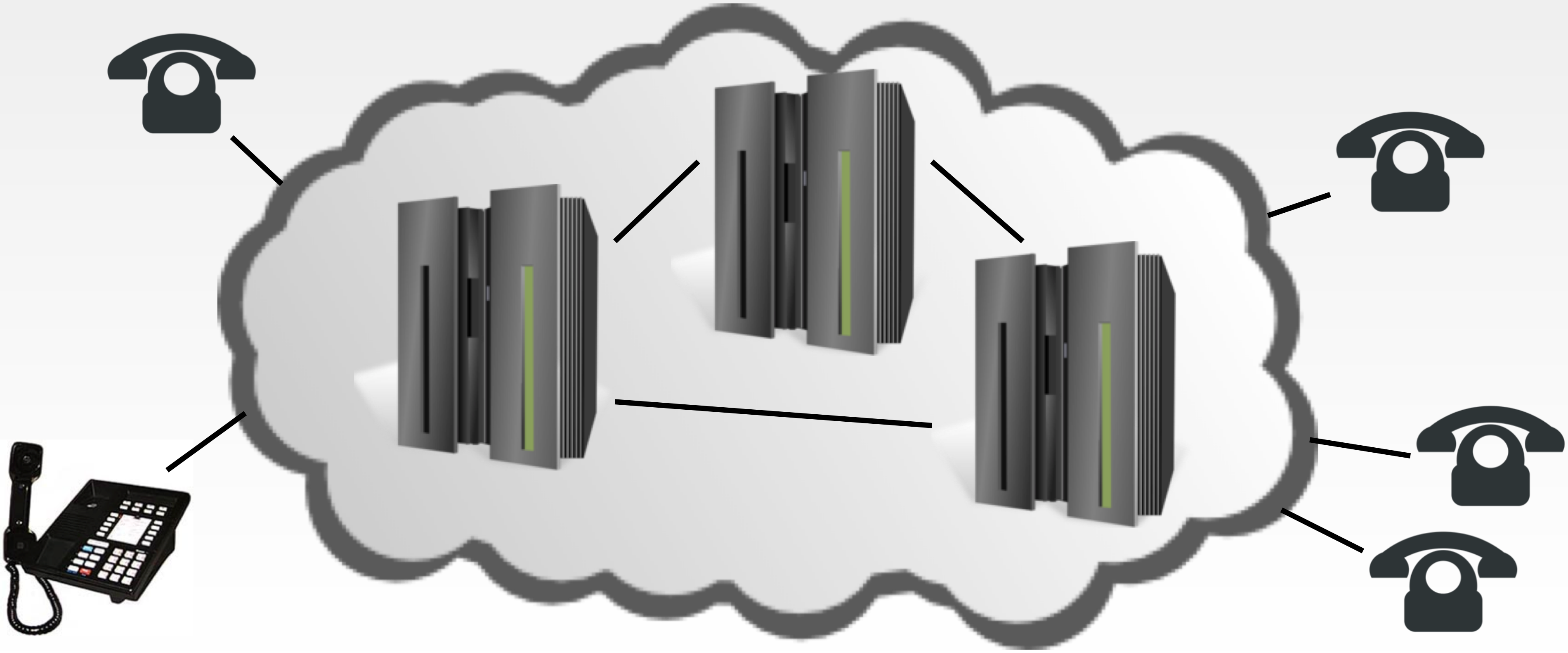
Web Browser

Web Server

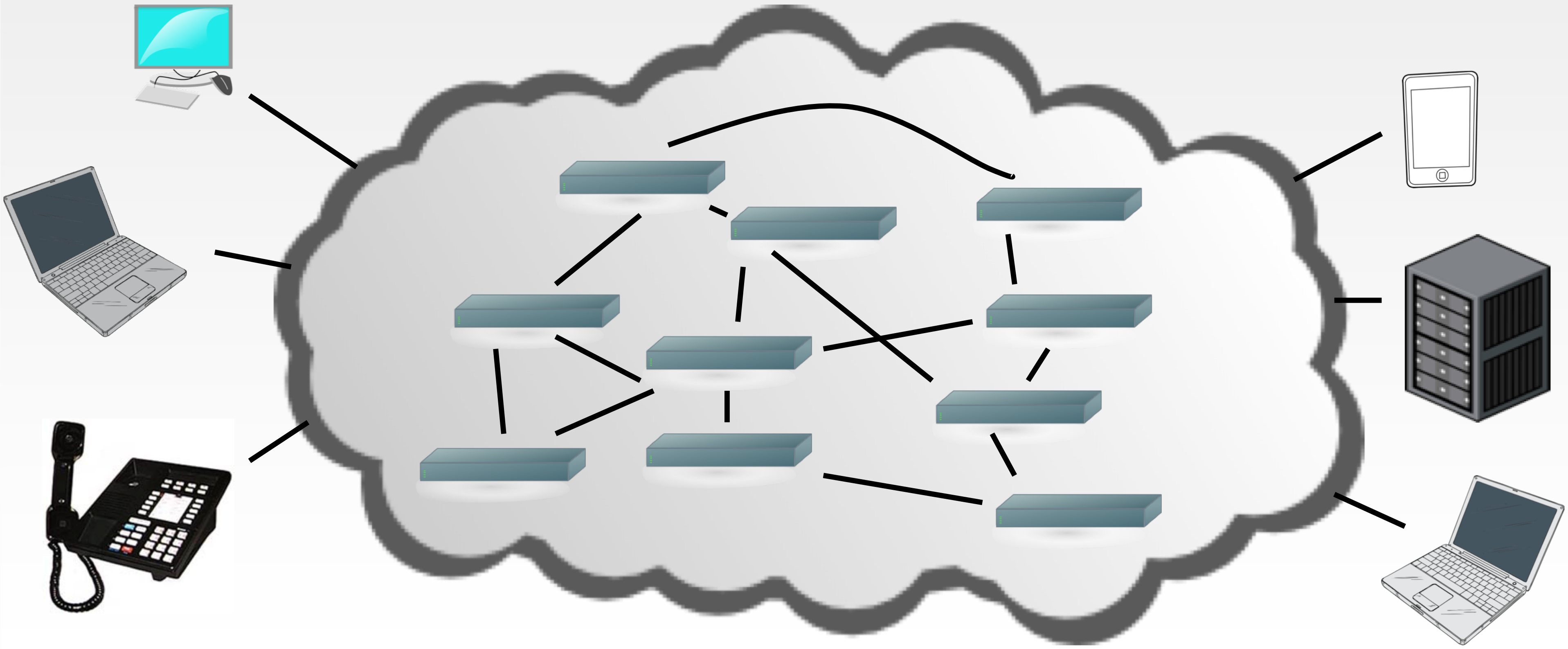
“Smart Hosts, Dumb Network”

- The phone network was built for dumb phones – nothing else was technically or economically feasible.
- All intelligence is in the network: conference calls, call forwarding, even many voice menus
- Internet routers are very dumb; all intelligence is in end systems
  - Consequence: *service* providers are not necessarily the same as *network* providers
  - Service provision is decoupled from physical location, and hence from jurisdiction
  - A person’s mail provider may be in another country

The Phone Network:
A Few Large Switches, Serving Phones



The Internet:
Many Routers, Very Many Types of Devices



Circuit Switching versus Packet Switching

- Circuits: traditional telephony model
- Path through the network selected at “call setup time”
  - Very small number of call setups; process can be heavyweight
- Each “phone switch” needs to know the *destination* of the call, not the source; return traffic takes the reverse path

- Packets: Internet model
- Every “packet” – a fragment of a message – is routed independently
  - No call setup
  - Routing must be very, very fast; it’s done for each packet
- Robustness: if a “router” fails, packets can take a different path
- Every packet must have a source and destination address, to enable replies
- Reply traffic may take a very different path

Many More Parties

- On the phone network, the central core knows everything
- On the Internet, the core knows little except for the endpoint addresses—and services can be provided anywhere
- Service providers can be and often are in different jurisdictions, where you may not have effective legal access (e.g., mail.ru)

IP Addresses

- A user types a name such as www.dni.gov.
- The *Domain Name System (DNS)* translates that to an *Internet Protocol (IP) Address* such as 23.213.38.42
  - IP addresses are four bytes long; each of those numbers is in the range 0-255
  - www.dni.gov actually uses a CDN, so every querier gets a different answer
  - (DNS resolution is complicated and involves many parties_
- IP addresses are what appear in packets
- Routers talk to each other (via *Routing Protocols*) to learn where each IP address is

IP Addressing

- Roughly 4 billion possible IP addresses today—we've essentially run out
  - IPv6, a newer version of IP being deployed now, has many more addresses
- IP addresses are handed out in blocks to big ISPs. Big ISPs give pieces of their allocations to smaller ISPs or to end customers
- Unless you're a very large enterprise, the only way to get IP addresses is from your ISP – and if you switch ISPs, you have to renumber your computers
- There is no analog to "local number portability" on the Internet – and can't be; there's no time to do that many lookups

Address Space Assignment

- IP addresses are handed out by *Regional Internet Registries (RIRs)*, such as ARIN
- They get their addresses from ICANN, an international non-profit which gets its authority from the U.S. Department of Commerce – controversial abroad
- Addresses are allocated based on demonstrated short-term need and evidence of efficient use of previously-allocated addresses
- Addresses may not be sold, even as part of a bankruptcy, merger, or acquisition, except with ARIN's approval and in accordance with ARIN's policies
  - This assertion of authority has never been contested in court—and some have been transferred by order of a bankruptcy court
  - Some ISPs have (very valuable) pre-ARIN addresses, called "legacy space". Legacy address holders don't have to renumber when switching ISPs (among other advantages)

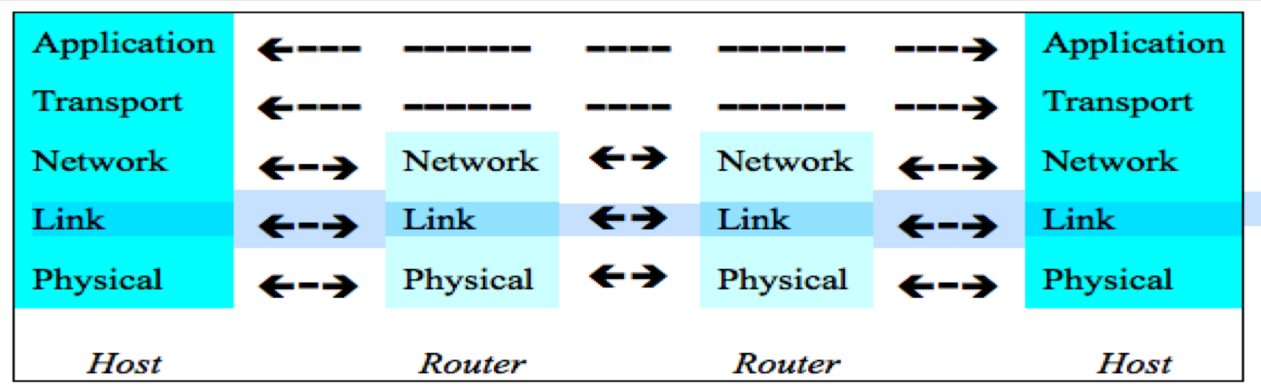
Implications of Address Space Allocation Policies

- That IP addresses are bound to providers *may* have antitrust implications—switching ISPs requires a lot of extra work to renumber computers
- Any party to the DNS name translation process—typically including the ISP, for most consumers—learns what sites are being contacted
  - Anyone who eavesdrops on the that traffic knows, too
- For a variety of reasons, including efficient use of IP addresses, ISPs always assign addresses hierarchically. This implies a tight, efficient relationship between IP addresses and location—important for online gambling, regional copyright licenses, jurisdiction

Port Numbers

- When one computer contacts another, is it trying to talk to a Web server or trying to send mail?
  - Remember that architecturally, all machines on the Internet are alike
  - It’s perfectly legal to run a Web server *and* a mail server on a single computer
- Packets contain not just an IP address but a *port number*
  - Port 25 is the mail server, port 80 is the Web server, 443 is encrypted Web, etc.
- If an IP address is like a street address, a port number is the room number in the building
  - Room 25 is the mail room, room 80 is library, etc.

The Network Stack



Application	←---	------	----	------	---→	Application
Transport	←---	------	----	------	---→	Transport
Network	←-→	Network	←→	Network	←-→	Network
Link	←-→	Link	←→	Link	←-→	Link
Physical	←-→	Physical	←→	Physical	←-→	Physical
Host		*Router*		*Router*		*Host*

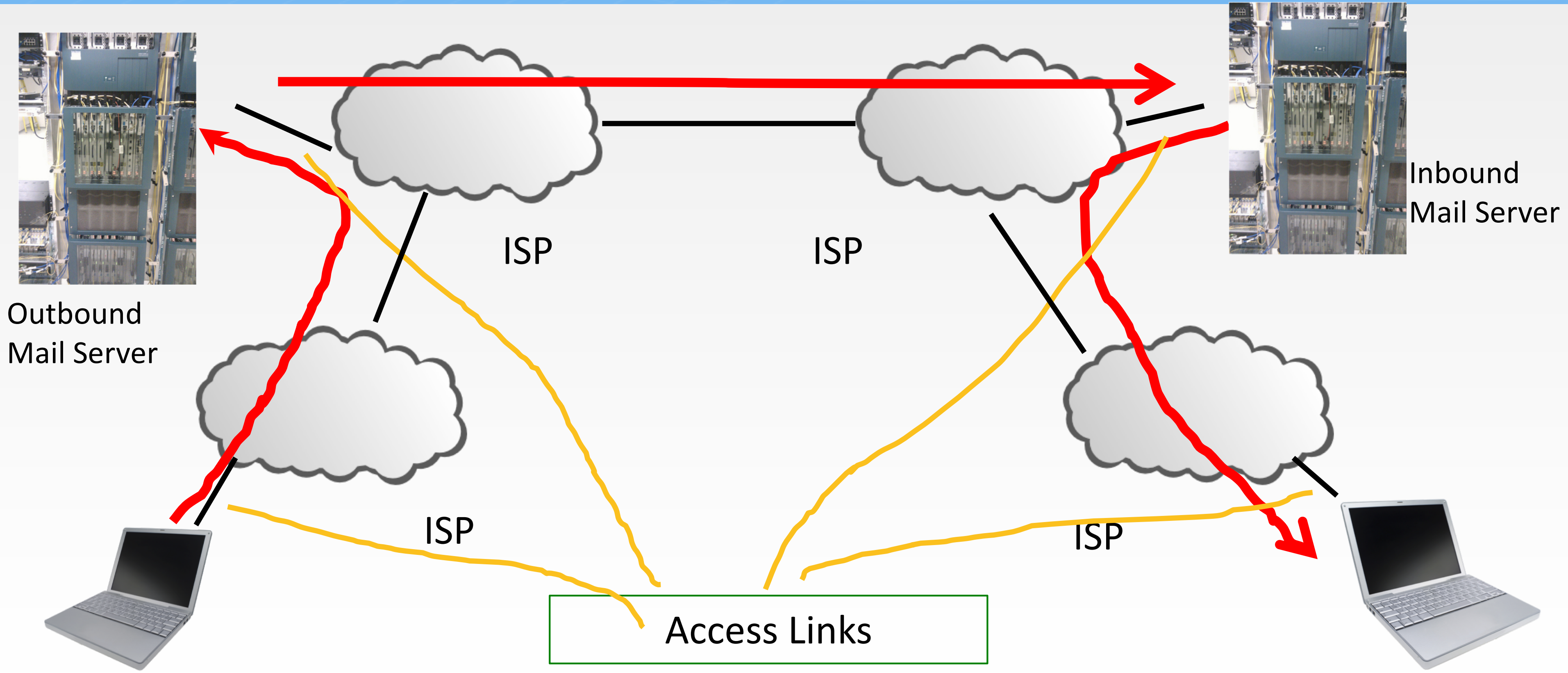
- The Internet uses a *layered* architecture
- Applications—email, web, etc.—are what we care about
- TCP (which has port numbers) *transports* the data; it is *end-to-end*
- IP (the *network layer*) is processed by every router along the path
- The *link layer* is things like WiFi, Ethernet, etc.

Are Port Numbers Third-Party Data?

- They’re not given to ISPs or other third parties (but IP addresses are very clearly accessible via the Pen/Trap statute)
- Well, under complicated circumstances they might be—but people very rarely know this
- ISPs sometimes examine them anyway, sometimes for consumers’ benefit and sometimes for their own reasons
- DoJ asserts that they are third-party data
- There’s no case law yet, nor any relevant cases that I know of

Email

Sending Email



Sending Myself Email—An SMTP Transcript

```
220 machshav.com ESMTP Exim 4.82 Tue, 11 Mar 2014 19:43:03 +0000
HELO eloi.cs.columbia.edu
250 machshav.com Hello eloi.cs.columbia.edu [2001:18d8:ffff:16:12dd:b1ff:feef:8868]
MAIL FROM:<smb@eloi.cs.columbia.edu>
250 OK
RCPT TO:<smb@machshav.com>
250 Accepted
DATA
354 Enter message, ending with "." on a line by itself
From: Barack Obama <president@whitehouse.gov>
To: <smb2132@columbia.edu>
Subject: Test

This is a test
.
250 OK id=1WNSaS-0001z5-1d
QUIT
221 machshav.com closing connection
```

Message

Conversation With A Third Party

```
220 machshav.com ESMTP Exim 4.82 Tue, 11 Mar 2014 19:43:03 +0000
HELO eloi.cs.columbia.edu
250 machshav.com Hello eloi.cs.columbia.edu [2001:18d8:ffff:16:12dd:b1ff:feef:8868]
MAIL FROM:<smb@eloi.cs.columbia.edu>
250 OK
RCPT TO:<smb@machshav.com>
250 Accepted
DATA
354 Enter message, ending with "." on a line by itself
[redacted]
.
250 OK id=1WNSaS-0001z5-1d
QUIT
221 machshav.com closing connection
```

Message

What the Recipient Sees

[redacted]

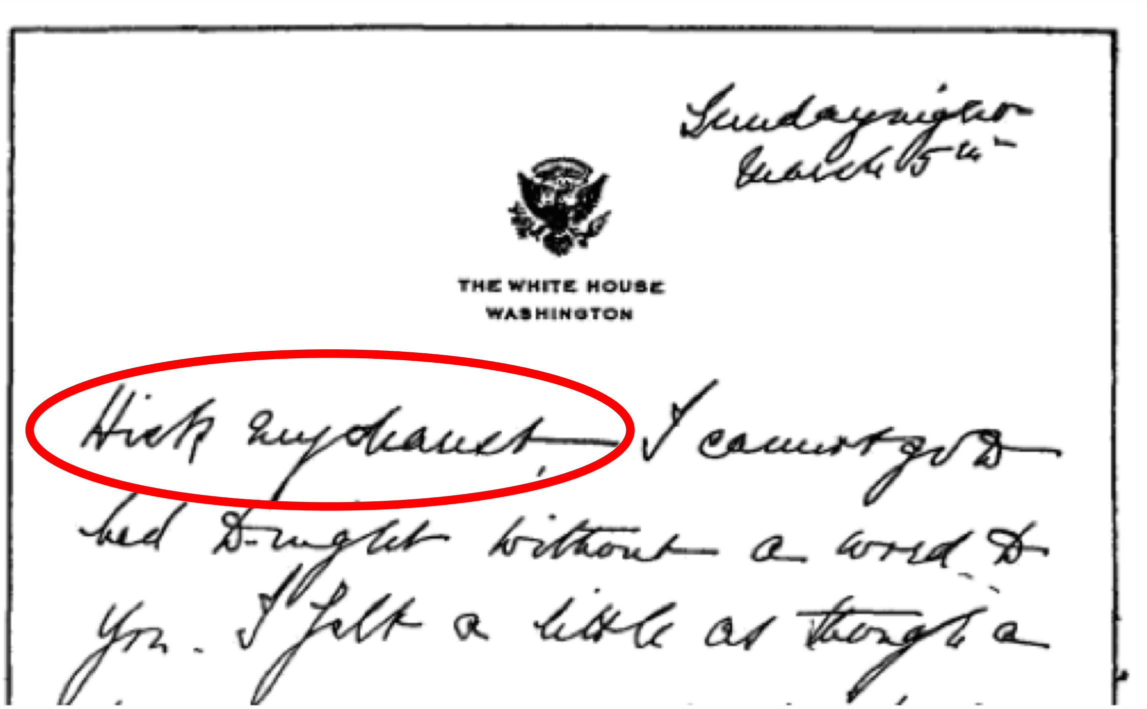
From: Barack Obama <president@whitehouse.gov>
To: <smb2132@columbia.edu>
Subject: Test

This is a test

Message

[redacted]

A Letter from Eleanor Roosevelt to Lorena Hickock (March 1933)



Sunday night
March 5th

THE WHITE HOUSE
WASHINGTON

Hick my dearest — I cannot go to bed tonight without a word to you. I felt a little as though a

It begins “Hick my dearest”.

(excerpt from Amazon.com)

Things to Note re the Third Party Doctrine

- The SMTP *envelope*—that's the technical term!—can have different information than the message headers
- Unlike the phone network, anyone can run their own mail servers
  - I personally run two, one personal and one professional
  - This complicates third party doctrine analysis
- The reality of email is far more complex than I've outlined here
  - Example: many people read their email via a Web browser—and the NSA has stated that even for them, picking out just the From/To information from a Webmail session is very difficult
- I haven't even begun to address server-resident email, virus scanning, spam filtering, and the like, let alone all of the other metadata that's present

Encryption on the Internet

Anything Can be Encrypted

- Links—though mostly used on WiFi
- Virtual Private Networks (VPNs)
- Simple connections (Web, email, etc.), generally via Transport Layer Security (TLS)
- Data, especially the body of email messages

VPNs

- Used by corporate employees for telecommuting or while traveling
  - Also used to connect multiple corporate locations
- Sometimes used to spoof location
  - Cover tracks
  - Fool geographic restrictions on content, e.g., streaming movies and music
- A recently published academic paper concluded that the NSA could cryptanalyze a lot of VPN sessions

TLS

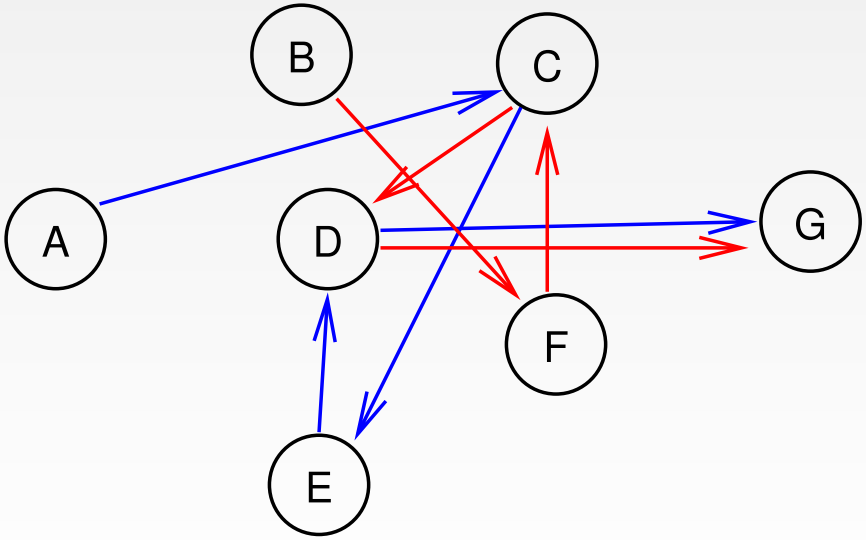
- Used for all secure Web traffic
- Widely (and increasingly) used when sending and retrieving email
  - But—TLS does not protect email “at rest”, i.e., while on disk on the various servers
- Used for many other point-to-point connections, e.g., Dropbox
- Older versions of TLS have cryptographic weaknesses; these are (believed to be) fixed in the newest versions
- The most common implementations of TLS have a long history of serious security flaws

Email Encryption

- Two different standards, S/MIME and PGP
  - S/MIME is widely supported—but rarely used
  - PGP requires less infrastructure support, and hence is used by enthusiasts
- Protects email at rest—but hinders searching
- Does not protect email headers or other metadata
- Both are generally very hard to use correctly

Tor: The Onion Router

- Computer A picks a sequence of Tor relays (C→E→D)
  - D is the exit node, and passes the traffic to destination host G
  - All of these hops are encrypted
- B picks relays F→C→D
  - G can't tell which is from A and which from B
- Neither can anyone else monitoring G's traffic
- Many use Tor for anonymity: police, human rights workers, spies—and criminals (e.g., Ross Ulbricht of Silk Road fame)
- Mental model: nested, sealed envelopes



Encryption

- Modern encryption algorithms, if used correctly, are *extremely* hard to break
- As a consequence, it is extremely hard to trace Tor connections or figure out the real origin of VPNed traffic
- The FBI and other law enforcement agencies have complained that they are “going dark” and want a legislative solution to provide for “exceptional access”
  - Most cryptographers think this is a bad idea

Cloud Computing

What’s a Cloud?

- A cloud is a traditional way to represent a network

- This “three-cloud network” picture is from 1982

- But—today “cloud” refers to computing services provided via the Internet by an outside party.

- (The modern usage seems to date to 1996: http://www.technologyreview.com/news/425970/who-coined-cloud-computing/)

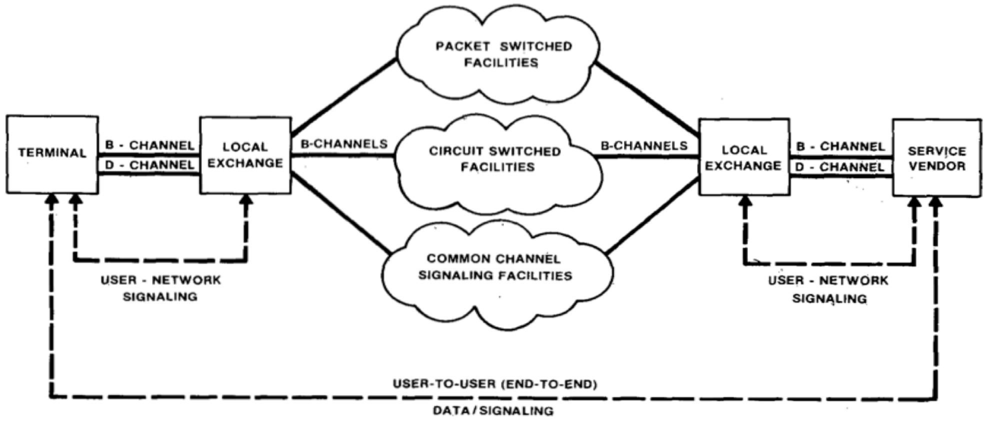


Fig. 7. The three-cloud network.

“Via the Internet”

- The service is not provided on-premises
- An Internet link is necessary
- This link provides an opportunity for interception, lawful or otherwise

“Outside Party”

- By definition, cloud services are provided by an outside party
  - Similar in spirit to the computing and time-sharing service bureaus, which date back to the 1960s

- *Not* the same as a company’s own remote computing facility
  - Organizations can have a “private cloud”, but the legal issues may be very different

Computing Services

- Many different types of services
  - Storage
  - Computing
  - Applications
  - Virtual machines
  - More

Storage

- Disk space in a remote location
- Easily shared (and outside the corporate firewall)
- Often replicated for reliability
  - Replicas can be on different power grids, earthquake zones, countries, continents, etc.
  - Data can be moved—or move “by itself”—to be closer to its users
- Expandable
- Someone else can worry about disk space, backups, security, and more
- Examples: Dropbox, Google Drive, Carbonite (for backups), Amazon S3
- Mental model: secure, self-storage warehouse

Computing

- Rent computing cycles as you need them
- Pay only for what you use
- Often used in conjunction with the provider’s cloud storage service
- Examples: Amazon EC2, Microsoft Azure, Google Cloud
  - Dropbox is a cloud service that uses a different provider’s cloud storage
- Mental model: calling up a temp agency for seasonal employees

Applications

- Provider runs particular applications for clients
- Common types: web sites, email services
- Less common types: shared word processing, payrolls
- Well-known providers: Google's Gmail and Docs, Microsoft's Outlook and Office 365, Dreamhost (web hosting)
- Mental model: engaging a contractor for specific tasks

Playing an Active Part: Google Docs

- Someone, using a Web browser, creates a document
  - Standard formatting buttons: font, italics or bold, copy and paste, etc.
- Others who have the proper authorization (sometimes just a special URL) can edit the document via their own Web browsers
- The changes made by one user show up *in real time* in all other users' browser windows
- In other words, Google is not just a passive repository; it is noticing changes and sending them out immediately

Virtual Machines

- Normal desktops: an *operating system* (e.g., Microsoft Windows) runs the computer; applications run on top of the operating system
- Virtual machines: a *hypervisor* running on a single computer emulates multiple real computers. A different operating system can run on each of these emulated computers—and each one is independent of the others and is protected from it
- Net effect: many computers that consume the space and power requirements of a single computer
- Mental model: rented office space

Location of Cloud Servers

- Responsiveness of and effective bandwidth to a server is limited by how far away it is
  - The problem is the speed of light—and not even Silicon Valley can overcome that limit!
  - It takes a *minimum* of a quarter-second to set up a secure connection from Washington to Paris, and twice that to New Delhi
- For performance reasons—and independent of political and legal considerations—large cloud providers therefore place server complexes in many places around the world
  - Also: take advantage of cheap power and cooling

Where is Data Stored?

- Modern email: on the server *and* on one or more devices
  - Users can’t easily tell what’s on their device (e.g., phone or laptop) versus what is retrieved from the server on demand
  - It differs for different devices at different times, and may depend on the user’s recent activity
  - What if the device and server are in different jurisdictions?
- (A bad fit for the assumed behavior model of Stored Communications Act)

Security and Privacy Issues

- Gmail: Google applications scan email and serve up appropriate ads
- Dropbox: uses Amazon S3 for actual storage; encrypts data so that Amazon can’t read it—but Dropbox can
- Spider Oak: data is encrypted with the user’s password; Spider Oak can’t read it
- Outlook.com: blocks file attachments that frequently contain viruses
- Many: check pictures for known child pornography
- Many: spam filtering

Compulsory Access to Email

- The Stored Communications Act requires search warrants for access to email less than 180 days old

- Older email is presumed to be abandoned and is accessible via a subpoena-like process
  - Widespread agreement that that provision violates the Fourth Amendment
  - But the government has argued that email is often voluntarily given to third party providers, so no search warrant is needed
  - In *United States v. Warshak*, the 6th Circuit strongly disagreed

Web Sites and Ads

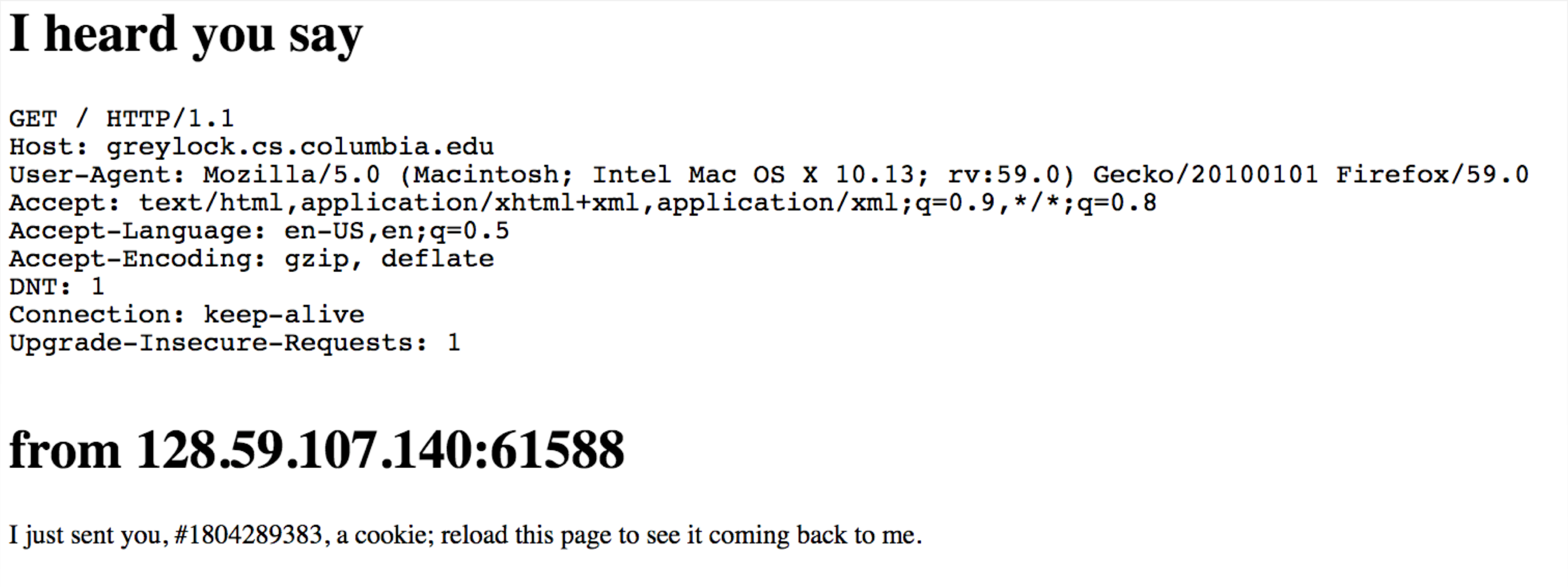
Web Pages

- Web pages are composed of many separate elements
- Images *always* are loaded from a separate URL
- There are also “frames”—web pages embedded in other web pages, again from separate URLs
- Many pages download JavaScript (small programs embedded in web pages) libraries from yet other URLs
- Web requests (HTTP—hypertext transfer protocol) generally contain the URL of the referring page
- In other words, for a typical web page *many* sites may know of the request

Cookies

- Web sites will tell you that cookies are small, harmless text files. That’s true, but…
- When you visit a site, it can set a cookie; your browser stores it on disk
- When you return to this site, your browser sends back that cookie
- Cookies are used for logins, site preferences, shopping carts, etc.
- They’re also used to track people around the web
- When you visit a page that includes other URLs, *each* site can set and receive cookies

The Initial Request



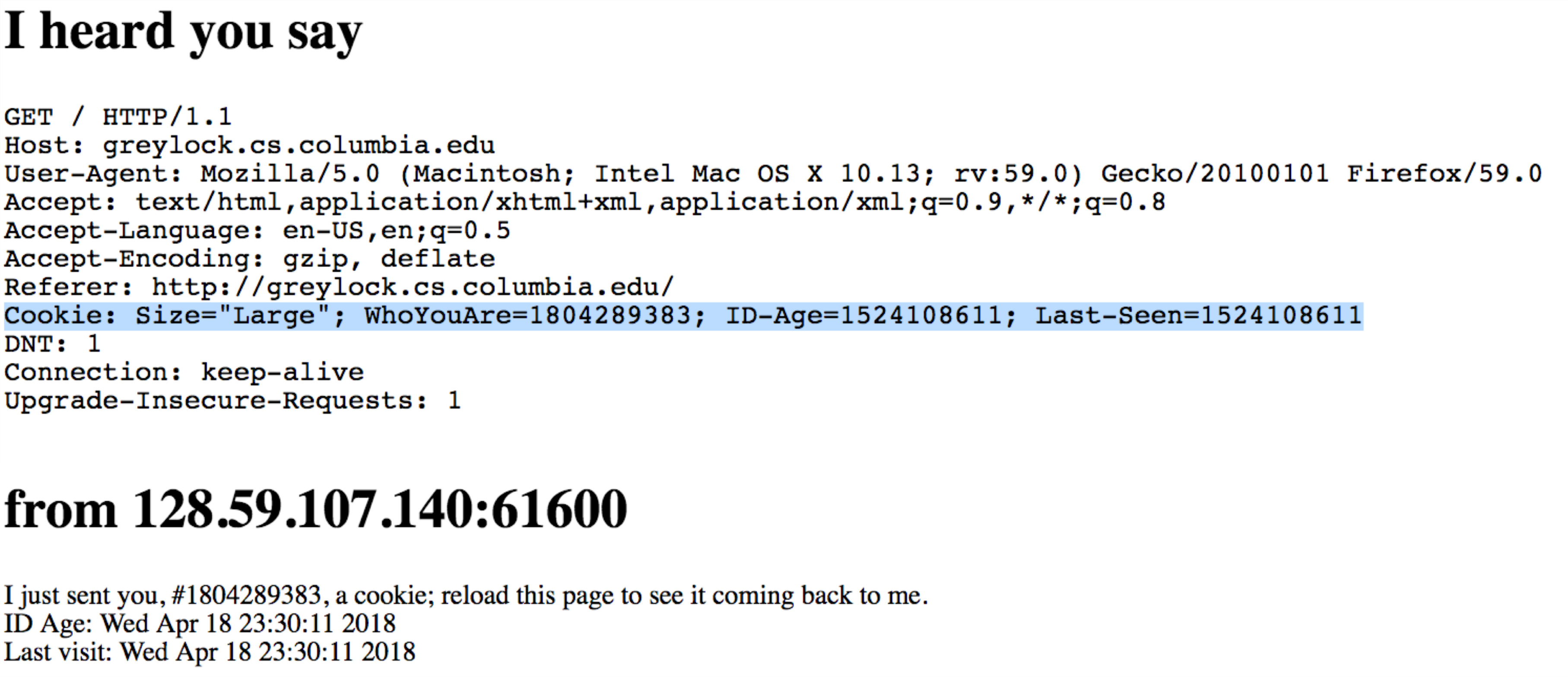
I heard you say

```
GET / HTTP/1.1
Host: greylock.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.13; rv:59.0) Gecko/20100101 Firefox/59.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
Upgrade-Insecure-Requests: 1
```

from 128.59.107.140:61588

I just sent you, #1804289383, a cookie; reload this page to see it coming back to me.

The Return Visit



I heard you say

```
GET / HTTP/1.1
Host: greylock.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.13; rv:59.0) Gecko/20100101 Firefox/59.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
Referer: http://greylock.cs.columbia.edu/
Cookie: Size="Large"; WhoYouAre=1804289383; ID-Age=1524108611; Last-Seen=1524108611
DNT: 1
Connection: keep-alive
Upgrade-Insecure-Requests: 1
```

from 128.59.107.140:61600

I just sent you, #1804289383, a cookie; reload this page to see it coming back to me.
ID Age: Wed Apr 18 23:30:11 2018
Last visit: Wed Apr 18 23:30:11 2018

Ad Networks

- Most ads are not served from the web sites you visit
- Instead, sites deal with ad brokers
- When you download a page, advertising image and frame requests point to ad brokers
  - The ad brokers get and set cookies
- They then send “redirect” commands, giving the URL of the actual ads for your browser to fetch and display
  - Again, the ad sites can get and set cookies

Privacy Implications

- You’re tracked around the web
- If you visit a site and click on an ad for, say, shoes, a cookie will be set by the ad network saying “this user is interested in shoes”
- When you visit a site that uses the same ad network, it will see that cookie—and show you more ads for shoes

Tracking

- Many web sites are involved in almost every visit to a web page
- Their log files are probably available as business records via the third-party doctrine
- Cookies are only accessible via a search warrant—but they (plus browser history) paint a very full picture of online activity
  - Note that they can disclose site logins, leading to more evidence
- Much of this tracking is dubious under the GDPR—but we'll have to see how web sites react
- These tracking-based ads provide the money that keeps the web operating
  - But whether or not tracking works is not that clear—the data is proprietary

Google, Facebook, et al.

Websites?

- Google and Facebook are web sites
- They’re also ad networks
- They collect massive amounts of information
- Their mobile apps let them collect even more, including your contacts and location
- And Facebook at least will merge information about online behavior with data about offline behavior

They know a tremendous amount about people, and
their machine learning algorithms let them intuit even more

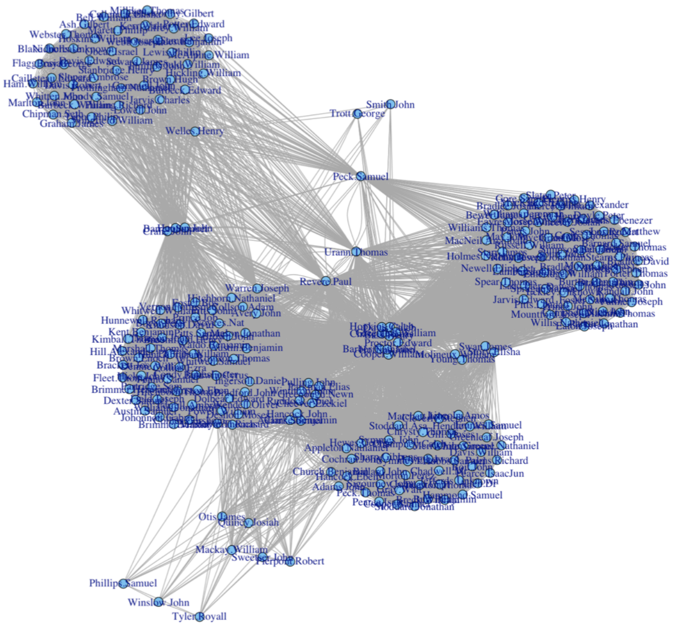
Ad Networks

- Together, Facebook and Google reap most of the profits (at least 65%) from online advertising
- Virtually all of the revenue growth has gone to these two companies?
- Why? Because they use their detailed knowledge about people for very precise targeting

Birds of a Feather

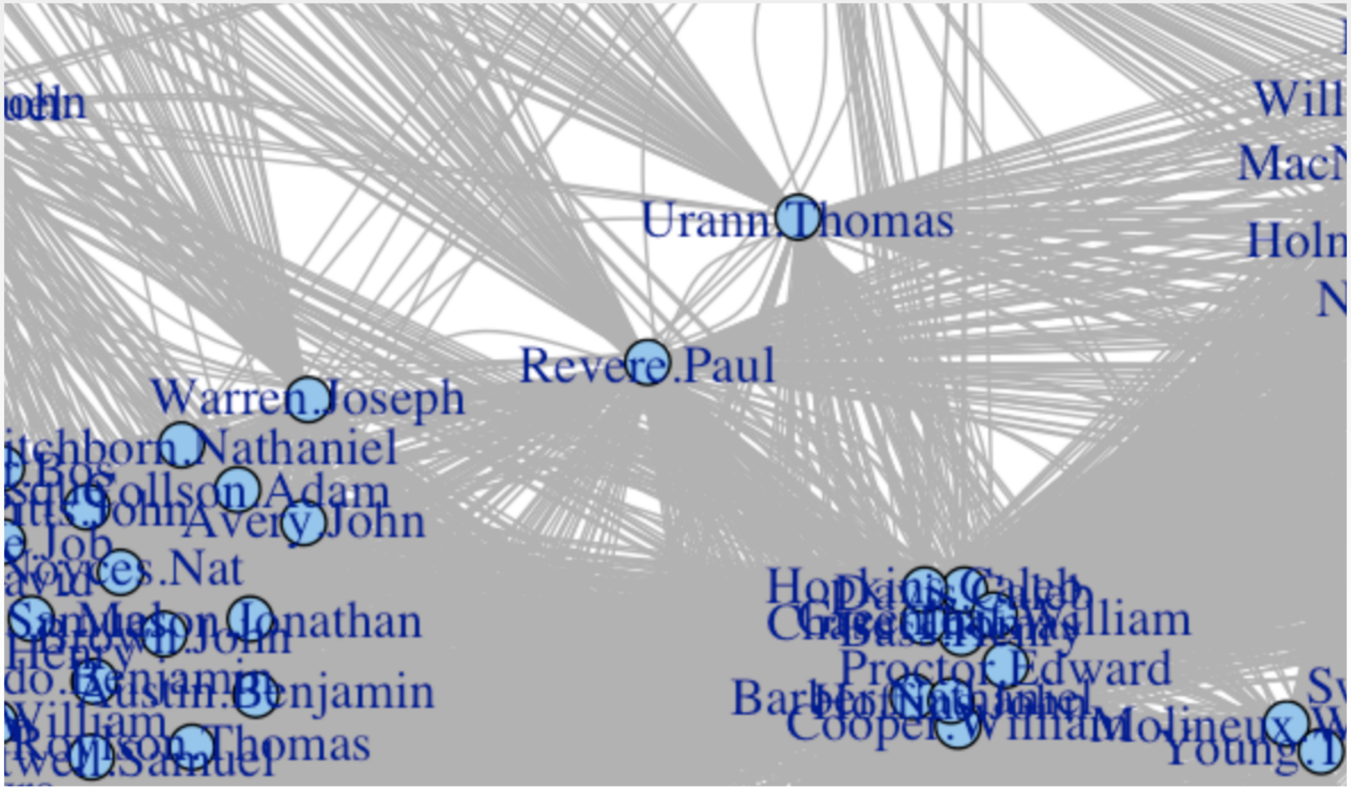
- Facebook, Twitter, and other social networks learn the *social graph*: who interacts with whom
- “Birds of a feather flock together” can be very true—and very revelatory
- Example: Some MIT undergraduates found that it was possible to predict people’s sexual orientation from whom their Facebook friends were

Using Metadata to Find Paul Revere



(https://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/)

And Right in the Middle…



Evidence

- In many cases, it has become routine to seize (often via subpoena) parties' Google and Facebook records
  - Divorce, disability, crimes, and more
  - Google records include search history, as well as email (via the Stored Communications Act)
- Don't forget the metadata
  - Communications patterns
  - Location
  - Timing of messages

Questions?