Privacy

Steven M. Bellovin
https://www.cs.columbia.edu/~smb

What is Privacy?

- Warren and Brandeis, 1890:

  > Recent inventions and business methods call attention to the next step which must be taken for the protection of the person, and for securing to the individual what Judge Cooley calls the right “to be let alone.”

- They proposed that a right to privacy, enforceable as a tort”, was consistent with existing law

- (Jewish tradition, going back 1800 years, holds that there is a right to privacy mentioned in the Bible: Numbers 24:5)

What is a Tort?

"A tort is an act or omission that gives rise to injury or harm to another and amounts to a civil wrong for which courts impose liability. In the context of torts, 'injury' describes the invasion of any legal right, whereas 'harm' describes a loss or detriment in fact that an individual suffers."

https://www.law.cornell.edu/wex/tort, quoting the *Restatement (Second) of Torts* § 7

A Civil Harm, Not a Regulation

- Warren and Brandeis did not propose explicit privacy dos and don'ts
- Rather, they relied on societal norms
- "An honest blunder, or a mistaken belief that no damage will result, may absolve the actor from moral blame, but the harm to others is still as great, and the actor's individual standards must give way in this area of the law to those of the public. *In other words, society may require of a person not to be awkward or a fool*." [Keeton et al.] (Emphasis added)

Privacy and Technology, 1890

Warren and Brandeis blamed “recent inventions and business methods” for privacy problems. And what were the ills?

> “Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that ‘what is whispered in the closet shall be proclaimed from the house-tops.’”
>
> … “the latest advances in photographic art have rendered it possible to take pictures surreptitiously”
>
> … “If you may not reproduce a woman's face photographically without her consent, how much less should be tolerated the reproduction of her face, her form, and her actions, by graphic descriptions colored to suit a gross and depraved imagination.”

Yes—photography! Gossip columns! (Imagine Warren and Brandeis in an era of social media.)

It Didn't Catch On

- Circa 1897, a flour company used a 14-year-old girl's photograph to advertise its products
- She sued—and lost: "others would have appreciated the compliment to their beauty"
- (In 1904, though, the judge who wrote that opinion complained about paparazzi stalking him when he was running for president…)
- New York passed a law prohibiting unauthorized use of someone's picture for advertising—and that law has been used against Facebook!

https://gizmodo.com/how-a-19th-century-teenager-sparked-a-battle-over-who-o-1829572319

Fast Forward: The 1960s

- Computers have really caught on in business
- People are starting to worry about computers and privacy
- Alan Westin publishes a famous book, *Privacy and Freedom*, summarizing the deliberations of a NYC Bar Association committee
  - A new concept: “data shadow”
- Congress holds many hearings

Westin:

- “Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.”

- “A central aspect of privacy is that individuals and organizations can determine for themselves which matters they want to keep private and which they are willing—or need—to reveal.”

- “[C]onsent to reveal information to a particular person or agency, for a particular purpose, is not consent for that information to be circulate to all or used for other purposes.”

In other words: privacy requires consent to release data, for particular uses.

Miller (Senate Hearing):

- “To insure the accuracy of the Center’s files, an individual should have access to any stored information concerning him and an opportunity to challenge its accuracy.”

- “Excessive reliance should not be placed on what too often is viewed as a universal solvent—the concept of consent. How much attention is the average citizen going to pay to a governmental form requesting consent to record or transmit information?”

In other words: We need transparency and accuracy, but people may not pay much attention to details.

The HEW Advisory Committee

- In 1973, the Department of Health, Education, and Welfare asked an advisory committee to study the problem of privacy
- Drawing on the work in the 1960s, the committee devised the *Fair Information Practice Principles (FIPP)*
- The FIPPs have become the basis for all privacy regulations, worldwide

The FIPPs

- No secret databases
- Individuals can learn what's stored about them and how it is used
- Individuals have the right to prevent secondary uses
- Individuals have the right to correct records about them
- Database operators must keep their systems secure

In other words: transparency, consent, correctness, security

Privacy Threats

Personally Identifiable Information

- Personally Identifiable Information (PII) underlies the FIPPs
- PII—names, addresses, social security numbers, etc.—*identifies* individuals
- In theory, if data does not contain PII, it does not violate privacy and hence is not covered by privacy laws
  - As we shall see, that's not always accurate

Data Brokers

- Data brokers: companies that collect and market large amounts of data on *everyone*
- They collect public records and buy data from lots of sources
  - Example: when you bring a car in to a mechanic, they sell your odometer reading
- They have *thousands* of data points on the average adult American
- This industry is almost completely unregulated

Unique Identifiers

- PII isn't always necessary for privacy violations
  - Amazon, Netflix, TiVo, etc., don't need to know your PII to build up a dossier on you
- What they need: a common identifier to link multiple transactions
  - iPhones now have "advertising identifiers", resettable at will and not linkable between apps
  - Apps used to use a constant phone ID
- Holy grail for marketers: a way to link mobile sessions, desktop/laptop sessions, and offline behavior

Secondary Uses

- Very often, there is not a privacy problem from the mere existence of data—collecting it is often necessary
  - Example: your doctor really needs your medical history
  - Google really needs to know where you are to give you directions
- Problems arise when the data is repurposed

Database Joins

- Suppose there are two databases with different information on people
- To merge the two, you need a common key in the different records
  - Names aren't great—many people have common names, and there are representation issues ("Steve" versus "Steven")
  - You can use secondary fields (address, date of birth, etc.) to help disambiguate
  - Some great ones: social security number, mobile phone number, credit card numbers
  - Unique identifiers!
- Conversely: you can protect privacy if you eliminate these database keys

Third Parties

- Many web sites (especially Google, Facebook, and advertisers) know something about what you do on other sites
- Many collect other information about you from third parties:
  - Washington Post: “We may receive information about you from publicly and commercially available sources…[and] from a social media site if you connect to the Services through that site.”
  - Wall Street Journal: “We may receive Other Information about you from third parties, including, … information about your interests, and information about your activities on other websites.”
- Plus: there is often third party content

Third-Party Content

- Most commercial sites include information from third parties
- These parties have their own privacy policies:
  - Washington Post: “Our Services may embed content from, or link to, third-party websites… that are outside of our control… This Privacy Policy does not govern these third party’s content or services, and we encourage you to review the privacy statements applicable to the third-party websites and services you use.
  - What are these third parties? How can you tell which they are?

Online Advertising

- Advertising is the business model of some of the biggest companies, most notably Google and Facebook
  - It has been called “the original sin of the Internet”
- Advertisers learn what you like, where you go on the net, and what you do
- This is generally with the cooperation of the hosting site

Cookies

- You are tracked on the web via “cookies” (and some similar techniques)
- A cookie is a string of arbitrary data that a web site can set and then read back the next time
  - Other sites can’t see that cookie
- Original purpose: login identifier, site preferences, shopping cart
- Today: how sites track return visits
- For fun: go to http://greylock.cs.columbia.edu (I’ll leave it up for a few days)

The Reason for Third-Party Content

- site1.com can't track site2.com
- Suppose both have Facebook "like" buttons
  - Those buttons are loaded from Facebook—so Facebook can set and receive cookies from each transaction
  - The "Referer:" HTTP header means that Facebook knows what page you came from each time
  - If you're logged in to Facebook in that browser, there's a Facebook login cookie
  - They can thus track you around the Web
- Google, Twitter, and others play the same games

Online Advertising

- Sites sell space to ad brokers; web pages reference ad broker sites (via IFRAMES) to load the ad
- The ad broker can invoke another ad broker via an HTTP Redirect; this new ad broker can do the same
- Eventually, an ad is loaded
- *But every ad site along the way can set and retrieve cookies, and thus track you*
- Google and Facebook control most of the online ad market, precisely because they're so good at tracking people

Location, Location, Location

- Location data is extremely revealing about people

  Justice Sotomayor: “Disclosed in [GPS] data . . . will be trips the indisputably private nature of which takes little imagination to conjure: trips to the psychiatrist, the plastic surgeon, the abortion clinic, the AIDS treatment center, the strip club, the criminal defense attorney, the by-the-hour motel, the union meeting, the mosque, synagogue or church, the gay bar and on and on.”

- It is thus extremely valuable to advertisers

- Mobile devices are great at knowing where they are

Ascertaining Location

- IP addresses are geographically assigned
  - There are commercial “IP geolocation” companies—and Google is better at it than they are
  - Requires no cooperation from users
- All smartphones have GPS receivers
- Cell tower triangulation
- WiFi or Bluetooth access point triangulation
  - Note: these latter two require sending information to a server—which therefore knows where you are before your phone does

Mobile Devices: Other Issues

- Mobile devices—today, phones—present huge privacy risks
- They're always with you
- They store a huge amount of data
- They have all sorts of sensors: microphones, cameras, accelerometers, etc.
- Even the US Supreme Court has recognized this

Other Metadata

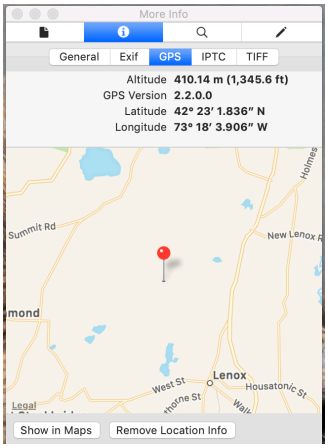
- Metadata—not the actual content of a conversation, but things that can show externally—can be very revealing
- IP addresses give away location
- Traffic analysis—who talks to whom—is revealing
  - Facebook knows the “social graph”
  - Some MIT students found that they could identify sexual orientation that way
- Pictures and other file types contain lots of metadata

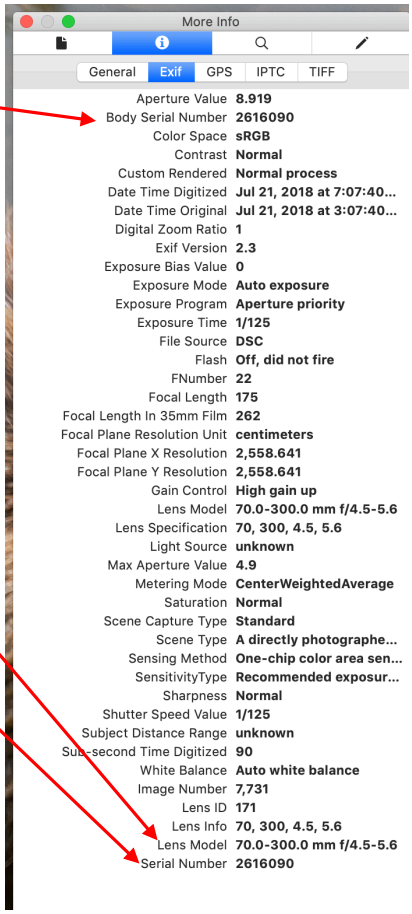
A Beaver

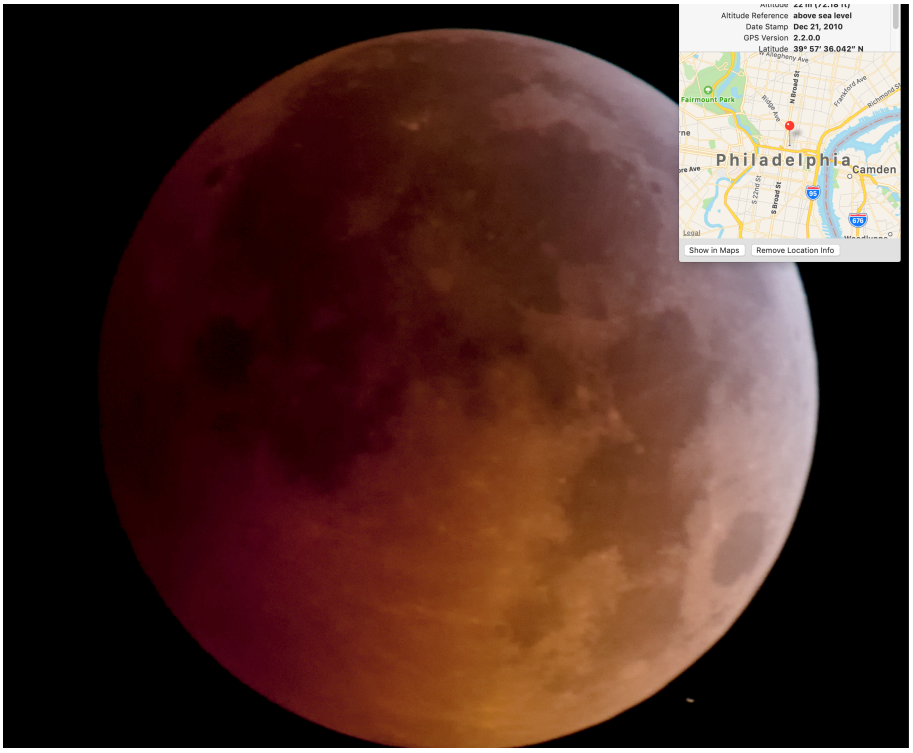
- Digital images contain EXIF data
- Some of it is purely photographic, e.g., camera settings
- Some is far more revealing

Photo Metadata

- The camera's serial number
- The lens model—how expensive is it?
- The lens serial number
- GPS location

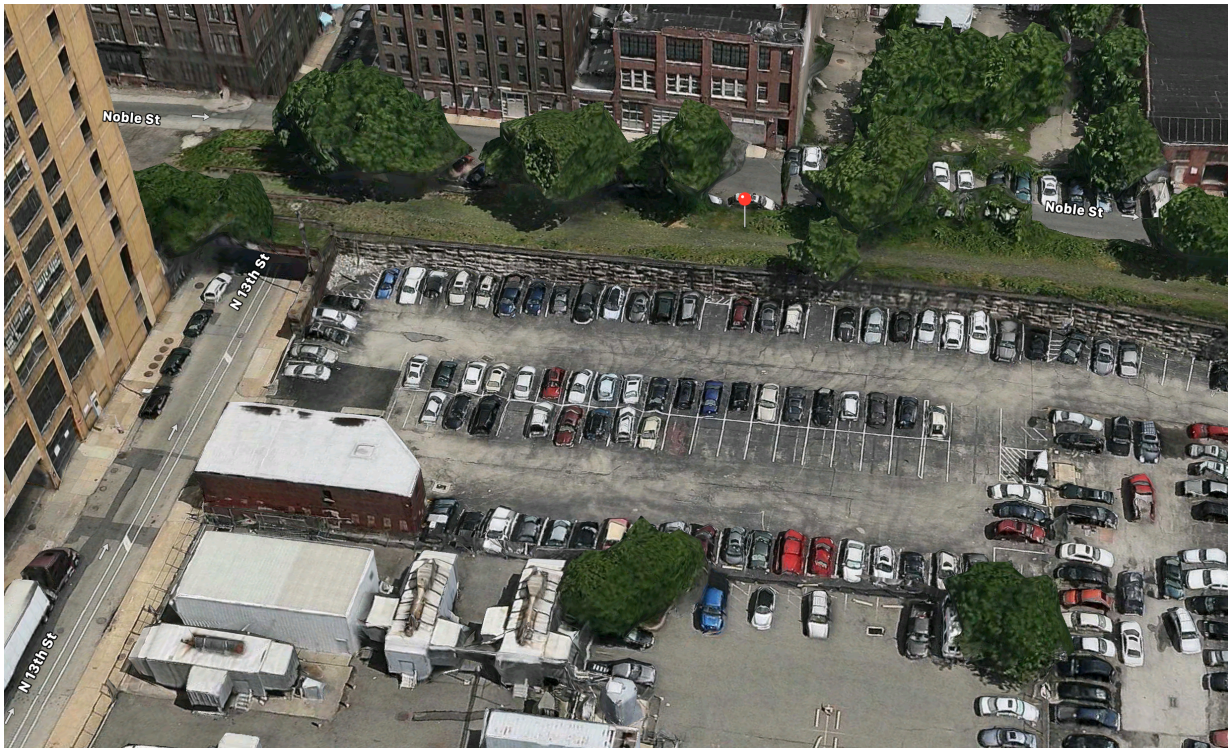






A Lunar Eclipse

(photo © Matt Blaze)



Where it was Taken

Bitcoin Metadata

- Bitcoin payments are to random-looking addresses: public keys
- But: all transactions, from one key to another, are recorded in the block chain
- You can trace payments from party to party—and if you ever learn someone's Bitcoin address, you can see their payment history *from public data*
- (There are newer cryptocurrencies that avoid this issue)

Machine Learning

- Very powerful data analysis technique used today
- Basic notion: feed in a pile of “training data”; machine learning algorithm can find patterns
- Possible to “know” things not known before
- Potentially very invasive of privacy

Types of Machine Learning

Supervised Machine Learning

- Manually label training data
- Algorithm looks at new data to see which labels it is closest to
- May not scale as well—but Google and others use user input to continuously retrain

Unsupervised Machine Learning

- Start without preconceived notions of what you have
- ML algorithm finds categories, items that cluster closer together
- New data matched against these categories

Privacy Invasion: Target

- Target learned that shoppers are creatures of habit: they tend to go to the same stores
- There are a few times when people switch stores, including during pregnancy
- Enter ML:
  - Identify new parents by sudden purchases of, e.g., diapers
  - Feed in their previous purchase history; use ML to identify leading indicators
  - Advertise baby products to other people who match that pattern
- They spotted a pregnant teenage girl before her parents knew…

Philosophical Questions

- Is data produced by an ML algorithm "known"?
  - "My TiVo thinks I'm gay"
- From a privacy perspective, is an incorrect guess better or worse than a correct guess?
- What if ML output is stolen in a data breach? Is this as bad as if authoritative data is stolen?
  - Note: a fair amount of nominally authoritative data has some errors
- (Note: many other systems issues with ML, e.g., biased training data will lead to biased output.)

Legal Background

Notice and Consent

- In practice, the FIPPs (and, in effect, US companies' privacy policies) rely on *notice and consent*
- That is, users are told what can happen; they then consent, implicitly or explicitly, to this
- But: do users actually know what can happen to their data? Can they grant consent to something they don't know about?

US Privacy Legal History

- 1970: The Fair Credit Reporting Act—much of the FIPPs (which hadn't been set forth yet…), plus some usage restrictions, e.g., a time limit on how long bankruptcies could be used for credit decisions
- 1974: The Privacy Act—effectively, the FIPPs, but *only* for the US government; the private sector wasn't covered. (Tidbit: government abuses during Watergate were one of the reasons the bill passed very quickly.)
- 1974: Family Educational Rights and Privacy Act (FERPA)—effectively, the FIPPs, but for students
- 1996: Health Insurance Portability and Accountability Act (HIPAA)—effectively, the FIPPs, but for health data

Etc—many sector-specific laws

Privacy Rights in Europe

- 1950: The European Convention on Human Rights, Article 8:

  1. Everyone has the right to respect for his private and family life, his home and his correspondence.

  2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

- 1980: The OECD effectively suggested the FIPPs to member countries

- 1995: The Data Processing Directive—*which applied to the private sector*

- 2018: The General Data Privacy Regulation

The US versus Europe

- Broad US privacy law regulates the government; private sector regulation is industry-specific
- The EU (and most other developed countries) regulate the private sector, too
- To a first approximation, the FIPPs underlies *all* privacy regulations—with, ironically enough, the smallest scope in the US

Commercial Privacy in the US

- There are many sector-specific rules
  - Details matter—the definitions are precise and picky; check with a lawyer if you *may* be covered
  - Example: is Microsoft's health record service covered by HIPAA? For a while, it was unclear
  - Example: if a covered HIPAA entity engages an outside party, e.g., for billing, there *must* be a specific agreement on HIPAA compliance
- For everyone else: privacy policies

Privacy Policies

- Most US web sites post “privacy policies”
  - They’re legally required in California (California Online Privacy Protection Act), so they’re effectively universal
- A privacy policy says what the site may or may not do with your data

Privacy Policies: *Not* the FIPPs

- Privacy policies are not required to be FIPPs-compliant
- Many, in fact, are not
- Privacy policies are an *agreement* with the user
  - More precisely, they're *imposed* on users: "click-wrap licenses"

Do Privacy Policies Protect Privacy?

Spoiler: Probably not

Do Privacy Policies Protect Privacy?

- People don't read them
  - Chief Justice Roberts doesn't read them
  - To read all relevant policies would take the average user 40 minutes per day
- They're often vague and incomprehensible
- And they're more about privacy invasion than privacy protection

Google's Privacy Policy

- They collect (among other things) unique identifiers, phone numbers, search terms, purchase activity, people you communicate with, activity on some third-party sites, location, etc.
- They use it for personalization, advertising, security, etc.
- But: they're better than most at letting you see your data, and you can delete at least some of it *if* you're logged in

Data Breach Notification

- Data breach: “The term ‘data breach’ means the loss, theft, or other unauthorized access… to data containing sensitive personal information, in electronic or printed form, that results in the potential compromise of the confidentiality or integrity of the data.”

- All 50 states have data breach notification laws—individuals must be notified if their data is compromised

- Note: generally tied to leakage of PII; must have enough information for identity theft or unusually sensitive data (e.g., health information) is leaked

Children's Online Privacy Protection Act (COPPA)

- Protects the privacy of children under 13
- Applies to web sites or online services (including mobile apps and connected toys) that are aimed at children or where the site has "actual knowledge" of users' ages
- Like the FIPPs, but with parents having control
  - In particular, parents must consent to collection of information about their children
- Protected personal information includes not just PII but also screen name, unique identifier, location, the child's image or voice, etc.
- Enforced by the FTC

Enforcing Privacy Policies

- A privacy policy is, in effect, a contract
- Who enforces it?
- In the US, it's often the Federal Trade Commission: the FTC
  - (Some state Attorneys-General will also enforce them)

The FTC

- The FTC can act against “unfair or deceptive acts or practices” that “causes or is likely to cause substantial injury to consumers.”
- If a company lies in its privacy policy, that is *a priori* deceptive, and is probably unfair competition as well
- Security failures due to carelessness can fall into this category, too
- But: the standards are often vague, and fines are quite low

Injury

- What is the actual harm from an invasion of privacy?
- If certain sensitive information is taken—credit card numbers, health data, etc.—that is clearly harm
- But: "trivial, speculative, emotional, and 'other more subjective types of harm' are usually not considered substantial for unfairness purposes"
  - But it may still fall under the "deceptive" clause
- Defining harm beyond the "creepiness factor" is hard!

Privacy Policies: Anything Goes

- Privacy policies are nothing of the sort: they do not guarantee privacy
- They're more properly termed "data collection and use policies"
- A privacy policy that says "we collect everything we can, from everywhere we can, and we do whatever we want with your data" is perfectly legal under US law

A Bill by Sen. Wyden: A New Hope?

- Establish minimum privacy and cybersecurity standards.
- Issue steep fines (up to 4% of annual revenue), on the first offense for companies and 10-20 year criminal penalties for senior executives.
- Create a national Do Not Track system that lets consumers stop third-party companies from tracking them on the web by sharing data, selling data, or targeting advertisements based on their personal information. It permits companies to charge consumers who want to use their products and services, but don't want their information monetized.
- Give consumers a way to review what personal information a company has about them, learn with whom it has been shared or sold, and to challenge inaccuracies in it.
- Hire 175 more FTC staff to police the largely unregulated market for private data.
- Require companies to assess the algorithms that process consumer data to examine their impact on accuracy, fairness, bias, discrimination, privacy, and security.

The GDPR

- Basically, the FIPPs with big teeth
- Applies to companies that operate in the EU *or* that collect data on EU residents
- Data breach notification
- Also includes the “right to be forgotten”
- Penalties for some infringements can be as high as €20 million or 4% of worldwide revenues
- Some non-compliant US sites have blocked access from the EU

The Right to be Forgotten

- Suppose that many years ago, someone has committed a crime, and that there are news stories about it
- Google never forgets—searches for that person's name *today* find the old news stories
- In 2014, the European Court of Justice ruled that this was an invasion of privacy
- Google was ordered to de-index those stories, but the news organizations were not forced to delete the articles
- In the US, the First Amendment (freedom of speech and the press) blocks any such ruling

The Right to be Forgotten: Issues

- Geographic scope: can a European court order Google to take down links from, say, their US sites?
  - French courts have issued such an order
- Is this censorship of truthful news items?
  - Is this American “constitutional fundamentalism”?
- But: People can be rehabilitated

Other Cross-Border Issues

- There are no boundaries on the Internet—which countries' laws should apply?
  - There is IP geolocation, but it can be inaccurate
- Where is cloud data stored? What if it's replicated?
- Sometimes there are national boundaries, e.g., the Great Firewall of China
- How do you reconcile conflicting aspects of privacy, e.g., anonymity (protected by US law) and improper disclosure of private facts?
- EU law often demands encryption to protect privacy; China often doesn't want it

Technical Aspects

Security and Privacy

- Security is *essential* for privacy
  - Privacy is the right to control where your information goes—if a hacker steals it, you've lost that ability
- But you need much more

Security Features for Privacy

- You need intrusion detection, to know when your defenses have failed
- You need exfiltration detection, to know when someone is stealing data
  - If nothing else, look at traffic volumes and destinations, and try to spot anomalies
- You need proper architecture

Security Architecture for Privacy

- Data breaches are more serious if PII is leaked together with sensitive data
  - Store the PII in a separate database
- Breach penalties are per person harmed
  - Make all accesses go through an API that logs everything
  - Keep the logs somewhere safe
- In other words: use internal privilege separation within your system

Privacy By Design

- Concept due to Ann Cavoukian, former Ontario Privacy Commissioner
- Basic notion: design in privacy, rather than trusting to security and process
- Examples:
  - Don't collect data, especially PII, if you don't need it
  - To the extent possible, process sensitive data immediately and then discard the input
  - Minimize possible database linkages
  - Encrypt what you can
- However: the concept itself is not well-defined; many have criticized it as being too vague

Measuring Privacy

- Two well-known techniques
  - *k*-anonymity
  - Differential privacy
- Not just academic—often used in the real world

K-Anonymity

- Devised by Latanya Sweeney and Pierangela Samarati (1998)
- Definition: in a dataset, a record for a given person is indistinguishable from those of at least *k* other people
- Achieved by suppressing certain fields
  - Obviously, must delete all database-matching keys
- One of the permissible techniques for releasing anonymized medical records under HIPAA

Differential Privacy

- Invented by Cynthia Dwork and Frank McSherry (2005)
- Intended to allow for statistical queries of data, e.g., average value of some field
- Add randomness to returned data
  - Too many statistical queries can permit reconstruction of the individual data items
  - Goal: average answers should be approximately the same if someone's record is removed
  - Tunable privacy versus accuracy parameter ε
- Used by Apple to protect personal data

Anonymization

- Purported solution to database privacy: replace PII with a random identifier
- The other information is still there, but there's no PII—meets the FIPPs requirements, right?
- Well, not really…

Deanonymization

- It is often possible to reverse anonymization
- General strategy: look for patterns in the data; match those against other data
- There are legal implications

The AOL Query Dataset

- In 2006, AOL released a dataset of 20,000,000 queries, with IP addresses and screen names replaced by random identifiers
- Two NY Times reporters managed to find one querier
  - Some queries were location-specific; others had a particular last name
  - Some queries were embarrassing or about medical conditions
- The person who released the data and his supervisor were fired; the CTO resigned

The Netflix Prize

- Netflix released 100,000,000 movie ratings
- They offered a $1,000,000 prize to anyone who could devise a better recommendation algorithm than the one they used
  - (That prize was claimed.)
- But IMDb had ratings by many of the same people
- It took remarkably few ratings to match Netflix users to IMDb users
- Despite this, Netflix started a second contest—but they were sued and investigated, and canceled the contest

Browser Features and Extensions

- Delete cookies on exit
- Block third-party tracking cookies
  - Safari and Firefox do this by default—but Chrome doesn't…
- The "Do Not Track" flag—but the semantics have never been defined, and no web sites honor it

Encryption

- What if data is stolen from disk?
- What if someone is eavesdropping on a network link?
  - Easy for WiFi; hard for the network backbone, though routing attacks can sometimes work
- Encrypt stuff!
- It's not that easy…

Who Has the Keys?

- Anyone who has the key can decrypt the data
- Encrypting communications is pretty easy
- Encrypting stored data is hard—the key has to be accessible for legitimate uses. Can the key be stolen, too?
- Perhaps store the key in an HSM (hardware security module)
- The attacker (probably) can't steal the key from the HSM—but can they send decryption requests to it?
- Log and monitor!

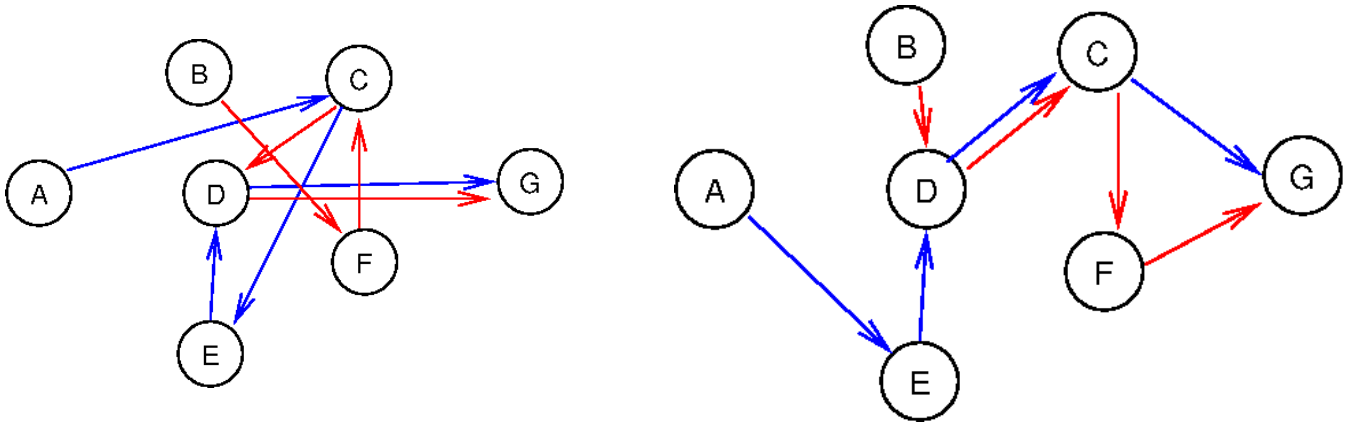
The Crypto Wars

- Law enforcement doesn't like encryption—it interferes with wiretaps, forensic analysis of seized devices, etc.
- They want a "golden key" to let them decrypt such data
  - Cryptographers call it a "back door"; the neutral term is "exceptional access"
  - Intelligence agencies don't insist much; their enemies won't comply no matter what
- Cryptographers generally insist that this can't be done safely

Tor: The Onion Router

- How can we hide IP addresses?
- Route traffic through random relay nodes
- Switch the path frequently
- Tor is safe against a limited adversary but vulnerable to a global adversary
- Original scheme from the US Naval Research Lab; the current Tor Project is run by the EFF

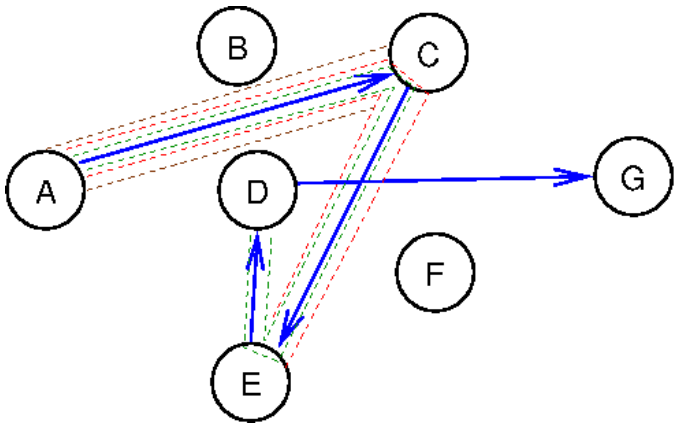
Two Different Paths to G



G cannot tell A from B

Tor Traffic is Encrypted

- Multiple layers of encryption when the traffic leaves A
- The first hop, C, strips off one layer—but C still can't read it
- E strips the next layer and forwards to D
- D strips the final Tor encryption—but A could use TLS to encrypt data all the way to G



Limits of Tor

- An adversary monitoring the links from A and to G can watch for correlated timings
- An active attacker can mark packets and watch for them
- Compromise Tor nodes, especially exit nodes
- Tor doesn’t protect higher-layer identifiers, e.g., logins
- Watch out for bugs in the Tor browser!

Operation PLAYPEN

- There was a child porn server operating over Tor
  - Tor has a mechanism for server privacy, too
- The FBI located the server and—with a warrant—seized it, operated it, and planted malware on it
- When users logged into this server, the malware was downloaded to their browsers—and it sent a non-Tor packet to the FBI saying “here I am!”
- That gave the FBI the IP address; they subpoenaed the customer name from the ISP

Where Are We?

Summary

- There are many threats to privacy
  - I've mostly talked about the private sector, but there can be government threats, too
- We have some defenses
- But it's very hard to live in the modern world, with things like smart phones and the web, and not leave a large data shadow

Questions?

(these slides at https://www.cs.columbia.edu/~smb/talks/cyberweek-gov_help.pdf)

Questions?

(these slides at https://www.cs.columbia.edu/~smb/talks/cyberweek-gov_help.pdf)