Privacy

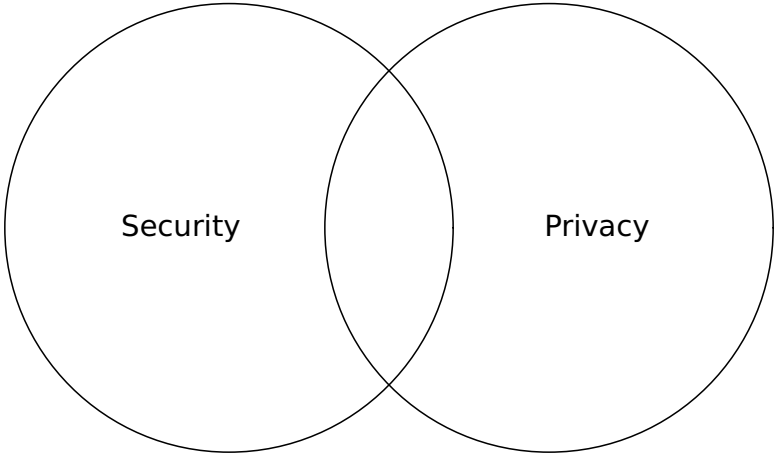
What is Privacy?

And why am I talking about it in a security class?

Privacy and Security Are Linked

- Many privacy breaches start as security breaches
- Security is a *requirement* for privacy
- Either sort of breach can be very costly
- Privacy breaches can lead to security failures

Security and Privacy Overlap



Source: Pollyanna Sanderson, Future of Privacy Forum

Equifax

- Equifax, a major data broker and credit bureau, was hacked; personal data on about 150,000,000 people was stolen
- ☞ (Many believe that a foreign intelligence agency was behind the hack)
- The FTC settlement could cost them up to $700 million
- Plus: up to $425 million in compensation and $1 billion to upgrade their data security
- But there was nothing of direct monetary value taken
- The root cause was a (series of) security and operational failures

Other Settlements

T-Mobile	$350 million to consumers
Capital One	$190 million
Zoom	$85 million

And many more

Definitions

anonymity The condition of an identity being unknown or concealed. (RFC 4949)

privacy The right of an entity (normally a person), acting in its own behalf, to determine the degree to which it will interact with its environment, including the degree to which the entity is willing to share its personal information with others. (RFC 4949)

confidentiality The property that data is not disclosed to system entities unless they have been authorized to know the data

Types of Privacy Violations

- Data (over?)collection for internal use, e.g., Google and Facebook
- Sale of data to other parties
- Theft of data due to a security problem

Security as a Privacy Requirement

	EU GDPR	CCPA	CA Ballot Initiative	WPA 2019	WPA 2020
Lawful Bases for Collection	Y	N	N	N	N
Privacy Policies	Y	Y	Y	Y	Y
Risk Assessments for High-Risk Activities	Y	N	N	Y	Y
Data Minimization	Y (strongest)	N	Y	N	Y
Purpose Limitation	Y (strongest)	N	Y	N	Y
Duty to Avoid Secondary Use	Y (strongest)	N	Y	N	Y
Reasonable Security	Y	Y	Y	Y	Y
Non-Discrimination	Y (Indirectly)*	Y	Y	N	Y

** The GDPR does not include an explicit provision stating that a data subject must not be discriminated against on the basis of their choices to exercise rights. However, it is implicit from other principles of the GDPR that individuals must be protected from discrimination on these grounds. (Article 5, Article 13, Article 22, and elements of "freely given" consent and fair processing).*

From https://fpf.org/wp-content/uploads/2020/02/fpf_comparison_of_wa_ssb-6281_to_gdpr__ccpa__cpra__and_2019_version_-_v1.0_feb_12_2020-1.pdf

Current and Proposed Laws

- *All* five require security
- Several other US states have passed privacy laws; others (and Congress) are considering them)
- An entity cannot determine the “degree to which the entity is willing to share its personal information with others” if some other parties can simply take it
- Security has been part of the requirements for privacy since the beginning

Compliance Obligations

- There are many privacy laws around the world
- The GDPR is the most famous, but all developed nations *except the US* have broad privacy laws
- The US has a variety of sector-specific laws (HIPAA, FERPA, Fair Credit Reporting Act, COPPA, etc) and state laws (e.g., CCPA, Illinois biometric act)
- You may need to do geolocation to figure out which laws apply to you
- But geolocation is itself a privacy risk!

Early Warnings

Senate committee hearing, 1967	Senator Long: But he could give that password to someone else, could he not? Dr. Piore: He can, and you find that some people do not protect their own password
Miller, 1969	“Another important security function that a privacy-oriented monitor program must perform is the identification of all users and terminals attempting to gain access to the files”
HEW committee, 1973	“Take reasonable precautions to protect data in the system from any anticipated threats or hazards to the security of the system”

FIPPs: Fair Information Principles and Practices

- First “code of fair information practices” developed in 1973 at HEW (Department of Health, Education, and Welfare)
- Basic rules for minimizing information collection, ensuring due process, protection against secret collection, provide security, ensure accountability
- Emphasize individual knowledge and consent
- Principles are broadly accepted (and form the basis of privacy law in the EU and many other places), but individual principles not implemented uniformly

Fair Information Principles and Practices (FIPP)

- Collection limitation
- Data quality
- Purpose specification
- Use limitation
- **Security**
- Openness/notice
- Individual participation
- Accountability

Note: these revolve around PII (personally identifiable information)

No Privacy, No Security

- Many password reset questions depend on private information
- Some sites, e.g., the IRS, use last year's data to authenticate new interactions
- Some private information can be used for blackmail or extortion
- An attacker can gain access to these sorts of information

So—How Do We Get Privacy?

Privacy is Old

- Jewish tradition from 1800 years ago finds a right to privacy in the Bible
- Semayne's Case, 1604: "The house of every one is to him as his castle and fortress."
- 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"
- In the 1960s, lawyers, academics, and Congress started worrying about privacy

Warren and Brandeis

- They worried about loss of privacy due to technological change
- Their threat: photography!
- “Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life”
- “the latest advances in photographic art have rendered it possible to take pictures surreptitiously”
- And new business models for newspapers: gossip columns!
- Imagine them in an era of social media. . .

(Relevant) Privacy Principles

- Do not collect more information than necessary
- Discard data when it is no longer necessary
- Avoid persistent identifiers
- Avoid linkable identifiers
- Avoid secondary uses
- Audit all uses

Overcollection

- Some information is collected but not needed
- Other information is needed for just a short time
- Get rid of it when feasible

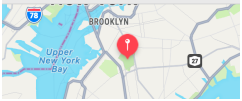
User Photos

- Digital photos contain a lot of metadata
- If you run a site that permits photo uploads, you can’t control what people send you
- But you can control what you do with it

Photos and Metadata

General Exif GPS IPTC TIFF

Altitude	34.02 m (111.61 ft)
GPS Version	2.2.0.0
Latitude	40° 39′ 34.71″ N
Longitude	73° 58′ 9.54″ W



General Exif GPS IPTC TIFF

Aperture Value	6.919
Body Serial Number	3028903
Color Space	sRGB
Contrast	Normal
Custom Rendered	Custom process
Date Time Digitized	Feb 8, 2023 at 1:03:15…
Date Time Original	Feb 8, 2023 at 1:03:15…
Exif Version	2.3.1
Exposure Bias Value	0
Exposure Mode	Manual exposure
Exposure Program	Manual
Exposure Time	1/100
File Source	DSC
Flash	No Flash
FNumber	11
Focal Length	800

Photo Metadata

- Location—obviously sensitive if it's someone's residence
- But photos can also contain the camera's and lens' serial numbers
- That can link photos taken by a given individual on different sites, with different usernames

Dealing with Photo Metadata

- You can’t help collecting it
- You can strip it before displaying to other users—Twitter and Facebook indeed do that (but are silent on what they do with such data internally)
- But—do you retain that data internally?
- (Photo metadata includes camera model—a useful clue to what someone will spend on photographic gear, and hence a clue for targeted advertising)
- (But what of photo-sharing sites like Flickr? Photographers often want to see the metadata of pictures they’re viewing)

IP Addresses

- Sites *must* know the IP address of users, and IP addresses convey location information
- It's always logged; these logs are necessary for operational reasons
- Location is also useful for advertising
- How long do you retain the logs? How long do you retain logs *linked to a user*?
- Is there an operational necessity to keep that more than a few weeks?
- Is there an audit-related reason to keep advertising-related data?

Persistent Identifiers

- A persistent identifier allows for long-term collection of information on a particular entity
- Examples: Web cookies, login name, email address, phone number, (US) Social Security Number
- Sometimes, an IP address is persistent

IPv6 Stateless Autoconfiguration

- IPv6 was designed so that hosts could get IP addresses without any infrastructure like a DHCP server
- For (presumed) uniqueness, use the interface's 48-bit MAC address as part of the IP address
- But—that becomes a persistent identifier, one that can even track laptops and other mobile devices across networks
- The IETF fixed it: hosts can generate random IPv6 addresses and check for uniqueness

Persistent Identifiers: Not PII!

- Many privacy laws focus on the presence of PII: Personally Identifiable Information
- PII: name, email address, street address, SSN
- Identifiers can be persistent, and hence potentially privacy-violative, without being PII
- Example: one of my Mastodon IDs is `@UrbanDinosaurs`—I don't use that as a login name anywhere else but there and Twitter, but my Mastodon activity under that login is traceable
- The account is not anonymous: "UrbanDinosaurs" is a *pseudonym*

PII Isn’t Needed For Privacy Violations

- NetFlix and Tivo know what you watch
- Google knows what you search for, and what you click on
- Amazon knows what you buy
- They don’t need your PII for any of that

Combining Sources

- Suppose you clear all cookies and do some Web searches
- Google builds up an anonymous profile of you
- You then log in—and Google combines the the anonymous history with its existing profile of you
- Or: Facebook will buy information from data brokers to combine with your online activity

Secondary Uses

- Per the FIPPs and the GDPR, when data is collected it must be for a specified purpose
- Today's US privacy policies are generally very vague about the purpose
- This permits secondary uses, which is where most privacy problems come from
- Glaring example: Facebook collected mobile phone numbers for login security, but then used them as persistent identifiers for user-matching
- Possible consequence: people will avoid 2FA, and lose out on its security benefits

Secondary Uses

- A driver's license to board a plane or enter a bar
- ☞ The swipe card readers some bars use for verification also record names, addresses, etc
- Digital rights management verifies that you've paid for the content—but it also tracks viewing habits
- The Medical Information Bureau tracks all health insurance claims in the US

Database Matching

- Some of the worst privacy problems occur when two or more independent datasets are joined
- Use persistent identifiers to match rows in tables
- Mobile phone numbers and SSNs are ideal for that—they very rarely change, and (especially for phone numbers) there is often an obvious legitimate reason for users to share them
- Combining multiple databases is the easiest way to find the real person who uses a given pseudonym

Collect or Retain: It's Not Easy

- Sometimes, data has to be collected
- Sometimes, there's a legitimate reason for retaining it
- But you may be able to achieve goals without compromising privacy

Case Study: TJX, 2005–2007

- TJX (owner of T.J. Maxx, Marshall's, and other stores) was hacked; credit card numbers and personal information was stolen
- This was a potential violation of Canadian privacy law; the Privacy Commissioner investigated
- The report is an interesting case study

Security Violations

¶80: WEP Several stores used WEP for WiFi security. WEP was known to be insecure but had not yet been replaced in their stores

¶16 From store LANs, intruders were able to attack internal machines, and moved laterally to a good place for information theft

¶22 Log files were deleted by the intruders, making it hard to track what they did

☞ In 1994, Bill Cheswick and I suggested keeping log files on a separate machine: “Hackers generally go after the log files before they do anything else, even before they plant their backdoors and Trojan horses. You’re much more likely to detect any successful intrusions if the log files are on the protected inside machine.”

Losses

¶77 Millions of credit card numbers might have been compromised
¶23 Names and addresses were taken
¶20 Driver's license numbers were taken

Reasons for Collection and Retention

¶63 Credit card number collection and limited retention is probably proper: “it may be reasonable to retain this personal information for the *length of time* [emphasis added] specified in the organizations’ contracts with financial institutions”

¶64: But. . . “with respect to retaining this information for ‘troubleshooting’ purposes, TJX/WMI has not presented a persuasive argument regarding the retention of this information for longer than 18 months”

Drivers' License Numbers

¶56 License numbers were collected and retained to deal with "return fraud"

¶58 But—after the hack, they decided to hash the license numbers; this meets their needs but protects privacy

Is Hashing License Numbers Secure?

- How large is the namespace?
- Many states have small namespaces—California, for example, uses a letter and 7 digits: 260,000,000 possibilities
- That's far too few—it's trivial to precalculate 260M hashes

Authentication and Privacy

- Usernames and email addresses can be persistent identifiers
- Biometrics are persistent identifiers
- Third-party single sign-on, e.g., “log in with your Google or Facebook account” is a privacy risk
- A single client-side identity certificate is a privacy risk

In privacy-sensitive environments, you must take this into account.

Credit Card Numbers and Privacy

- Stores track you by your credit card number
- Especially useful for stores with physical and online presence—transactions can be linked
- An old payment protocol intended to replace the need for stored credit card numbers online nevertheless included them, for precisely that reason

Side Channels

- Another privacy risk: side channels
- A side channel "is any communication channel that is incidental to another communication channel"
- Example: the timing of cryptographic operations can leak key bits
- Browsers leak *lots* of information, via fonts, languages, extensions, and more

Browsers

- Browsers leak *lots* of information
- To test your browser thoroughly, go to https://panopticlick.eff.org

I heard you say

```
GET / HTTP/1.1
Host: greylock.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15; rv:72.0) Gecko/20100101 Firefox/
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;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 128.59.23.26:63049

I just sent you, #846930886, a cookie; reload this page to see it coming back to me.

I heard you say

```
GET / HTTP/1.1
Host: greylock.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15; rv:72.0) Gecko/20100101 Firefox/72.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
Referer: http://greylock.cs.columbia.edu/
Cookie: WhoYouAre=846930886; ID-Age=1582059315; Last-Seen=1582059315
Upgrade-Insecure-Requests: 1
```

from 128.59.23.26:63227

I just sent you, #846930886, a cookie; reload this page to see it coming back to me.
ID Age: Tue Feb 18 15:55:15 2020
Last visit: Tue Feb 18 15:55:15 2020

Panopticlick: Firefox

Within our dataset of several hundred thousand visitors tested in the past 45 days, only **one in 2010.63 browsers have the same fingerprint as yours.**

Currently, we estimate that your browser has a fingerprint that conveys **10.97 bits of identifying information.**

The measurements we used to obtain this result are listed below. You can **read more about our methodology, statistical results, and some defenses against fingerprinting here**.

Browser Characteristic	bits of identifying information	one in x browsers have this value	value
User Agent	6.44	86.55	Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15; rv:72.0) Gecko/20100101 Firefox/72.0
HTTP_ACCEPT Headers	4.6	24.18	text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8 gzip, deflate, br en-US,en;q=0.5
Browser Plugin Details	3.0	7.99	no javascript
Time Zone	3.0	7.99	no javascript
Screen Size and Color Depth	3.0	7.99	no javascript

Panopticlick: Safari

Within our dataset of several hundred thousand visitors tested in the past 45 days, only **one in 16237.5 browsers have the same fingerprint as yours.**

Currently, we estimate that your browser has a fingerprint that conveys **13.99 bits of identifying information.**

The measurements we used to obtain this result are listed below. You can **read more about our methodology, statistical results, and some defenses against fingerprinting here.**

Browser Characteristic	bits of identifying information	one in x browsers have this value	value
User Agent	8.36	327.56	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_3) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/13.0.5 Safari/605.1.15
HTTP_ACCEPT Headers	4.66	25.32	text/html, */*; q=0.01 gzip, deflate, br en-us
Browser Plugin Details	6.43	86.4	Plugin 0: WebKit built-in PDF; ; ; (Portable Document Format; application/pdf; pdf) (Portable Document Format; text/pdf; pdf) (PostScript; application/postscript; ps).
Time Zone	3.17	8.97	300
Screen Size and Color Depth	5.09	34.03	2560x1440x24

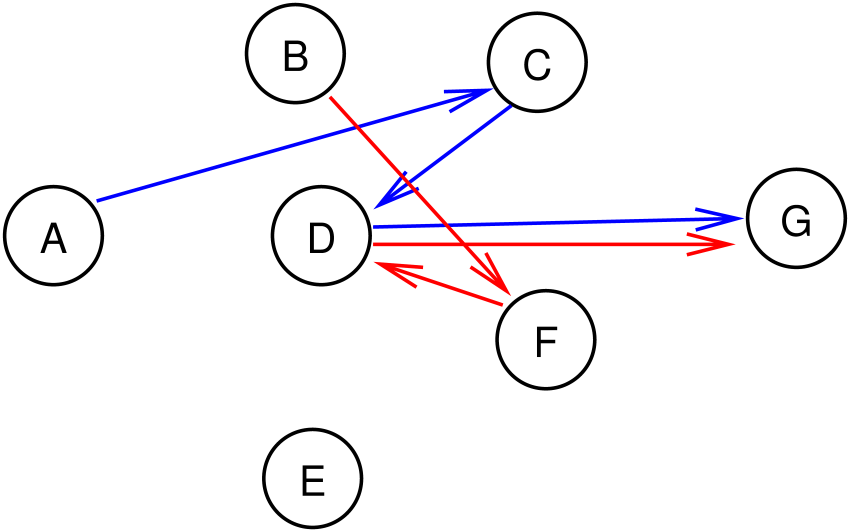
Tor

- Tor: The Onion Router
- A popular anonymity technology—blocks traffic analysis
- Originally developed at the US Naval Research Lab
- Picked up by the EFF; has received funding from the US State Department because of its use by dissidents and human rights workers
- Traffic routed via changing relay and exit nodes

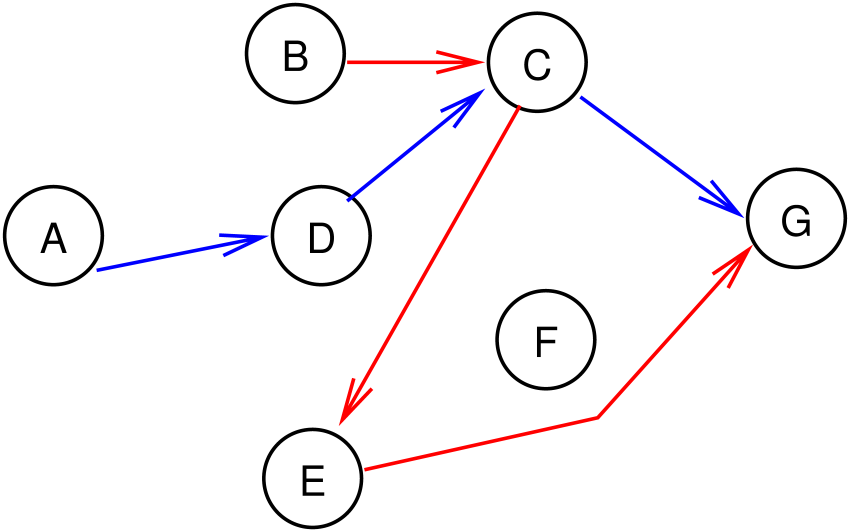
Onion Routing

- Pick a relay node
- Pick an exit node
- Send multiply encrypted traffic to the relay, thence to the exit, thence to the destination
- Not good against a “global adversary”—but real adversaries can’t see the whole Internet
- Caution: connections from Tor exit nodes are generally not from that site—and blocking it can affect many people

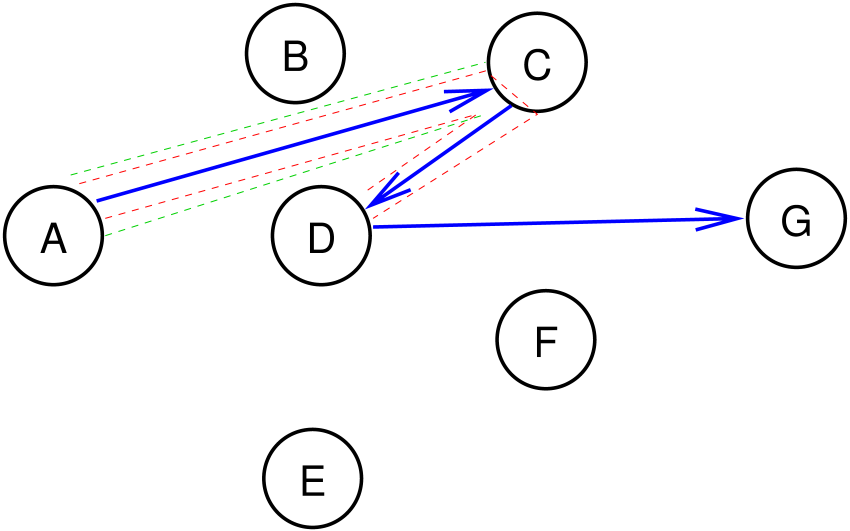
Onion Routing



Onion Routing



Onion Routing: Encryption



Tor Hidden Services

- Web servers can also set up listening posts on Tor
- They pick a bridge node and tell a directory server
- Clients going to a `.onion` URL use Tor to reach that node
- Tor hidden services now used by the NY Times, the BBC, Facebook, and many others

Privacy and Execution Environments

- Privacy violations can be seen as someone else gaining access to individuals' data
- Using data for one purpose when it was collected for a different purpose is a security violation by our definition: the reuse was not authorized by the individual
- Merging databases is another kind of reuse, but one with more serious consequences

Privacy and Security

- Protecting privacy is often a legal requirement
- You cannot protect privacy without strong security

Questions?

(Barred owl, Central Park, October 11, 2020)