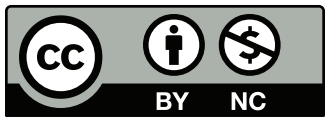
What is Privacy?



What is Privacy?

- Warren and Brandeis (1890): “the right to be let alone”
  (They were trying to find a legal rationale to protect privacy under then-existing statutes and principles.)
- FIPS PUB 41: “The right of an entity . . . 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”
- OSI: “The right of individuals to control or influence what information related to them may be collected and stored and by whom and to whom that information may be disclosed”

Using versus Gathering

- The primary concern is how information is *used*
- Obtaining information is often much less of a concern
- Note, though, that a lot of personal information is considered private even from one other person

Legal Foundations of Privacy

- Common law: “[T]he house of every one is to him as his castle and fortress.” Semayne’s Case, 5 C. Rep. 91a, 77 Eng. Rep. 194 (K.B. 1603)
- Doesn’t work as well in today’s interconnected world
- Information is collected, stored, analyzed

The Role of Computers

- Computers make mass storage (more) feasible
- (Punch card storage (1880s) started the process)
- Computers allow for rapid, sophisticated matching and correlation
- Computers can make inferences and predictions, and group people into categories

Inferences

- Amazon, Netflix, etc., try to predict what else you might like
- These algorithms work by correlation
- Often, they're right, but sometimes, they give odd results. . .

Amazon's Recommendation

Customers Who Bought This Item Also Bought

Page 1 of 20

Samsung Galaxy S4 Charger 2.1Amp 2-Port Adapter for Travel Home Wall with 3 feet Micro…

★★★★★ 22

$9.99 Prime

iphone 6s Full Screen Protector, PLESON® iphone 6 6s Edge to Edge Full Screen Cover [3D…

★★★★★ 41

$14.99 Prime

Soniworks Compatible (2-Pack) Replacement Facial Cleansing Brush Heads, designed for…

★★★★★ 106

$11.95 Prime

Hard Rhino Creatine Monohydrate Micronized 200 Mesh Powder, 125 Grams

★★★★★ 241

$10.01 Prime

Miss Jordan Salt and Pepper Grinder Set. Elegant Stainless Steel Salt and Pepper Grinder Set…

★★★★★ 56

$23.00 Prime

Why Violate Privacy?

- Thoughtlessness
- Efficiency, especially for marketing
- New markets (i.e., new location-based offerings)
- Public safety and national security

How Do We Lose Privacy?

- Voluntarily
- Compulsion
- Reuse of data

☞ *This sort of secondary use is the source of most privacy violations*

Voluntary Surrender of Data

- Social networking sites
- Purchases (Netflix, Amazon)
- Warranty cards

Compulsion

- Various interactions with governments (marriage, property purchases, etc.)
- Boarding an airplane
- “Contracts”—e.g., getting a credit card in exchange for information

Secondary Use

- We may not object—or object too much—to the initial collection of certain data
- Often, we benefit from the initial collection, and hence regard it as a fair trade
- When it is used for another purpose without our knowledge or consent, trouble often results

Example: Bars and Drivers' Licenses

- Many bars use swipe readers to verify that the proferred license is genuine
- (Better-grade fakes have mag stripe data anyway. . .)
- But—the readers copy the data: name, address, gender, etc.

What are the Privacy Violations?

- Using license data to establish age
- Using license data for marketing

Data on a Driver's License

- Primary purpose: certification that you are legally allowed to drive
- Primary purpose of picture: assurance that the bearer is indeed the license holder
- Demographic data: accountability in event of violations
- *Not* intended for proof of age, *not* intended as an airplane boarding credential

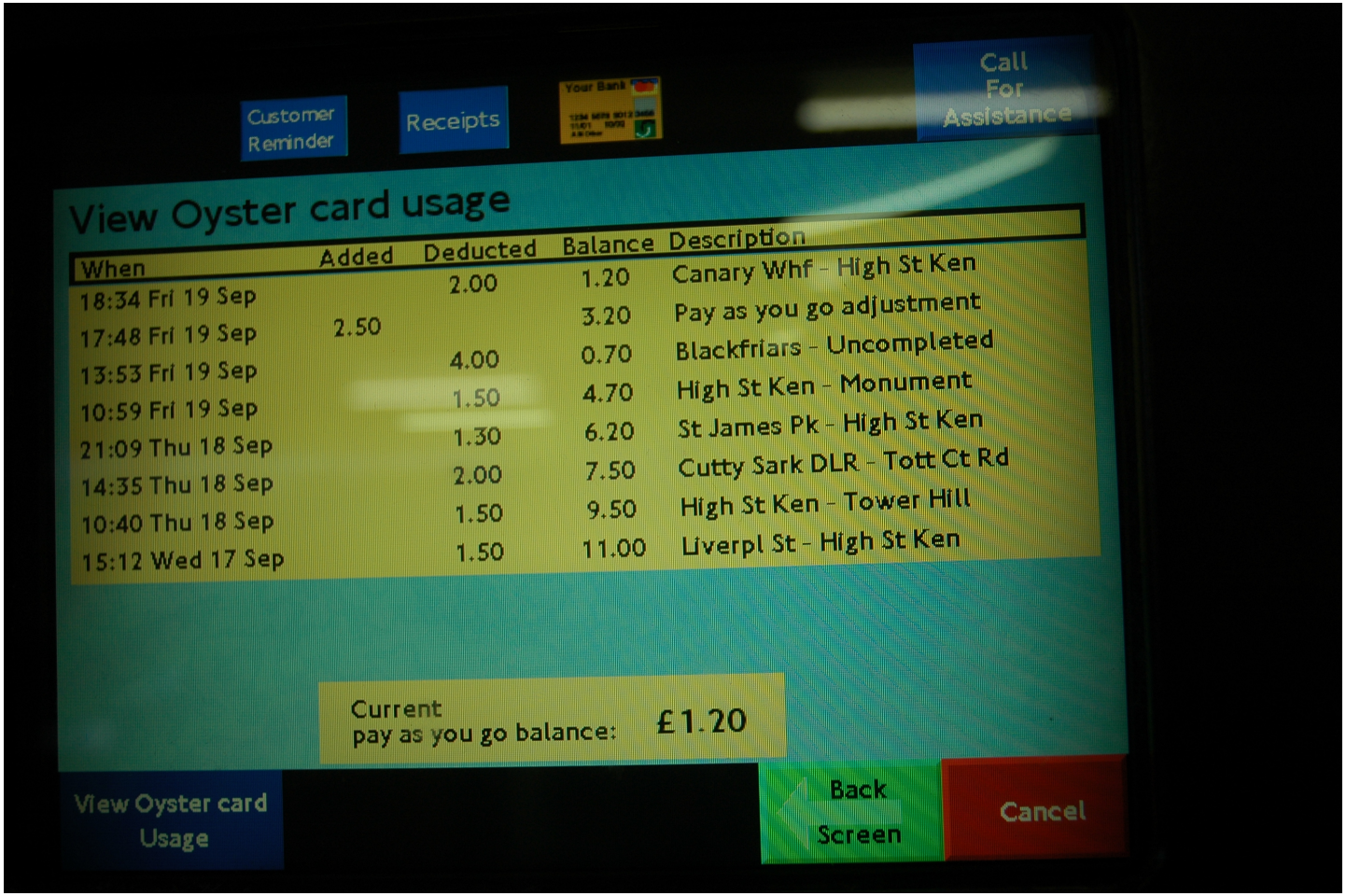
Age Verification

- Even if age verification is acceptable—and use of licenses for that is certainly accepted by the states—use of the additional data for marketing is not
- Resale of license data happens to be illegal, but not for that reason

Example: MetroCard

- Primary purpose: paying subway or bus fare
- But—the MTA retains your trip information
- This data can be and has been used for criminal and divorce cases

The London Oyster Card



Linkages

- Sometimes, items from two or more databases are linked
- Then possible to learn *much* more
- Prerequisite: common data item

Linkages: MetroCard

- How did you pay for your last MetroCard? Credit card?
- That links the MetroCard to a person
- Query: who boarded the subway at 116th and Broadway between 3:30 and 3:45 AM last Tuesday?
- In principle, at least, that question may be answerable

Deeper Linkages

- Correlate on patterns
- Example: assume a MetroCard is used infrequently, but at only two stops, Penn Station and 116th St
- Is there any one person who used a credit card to buy train (Amtrak, NJ Transit, LIRR) tickets on just those days?
- (Note: I have no idea if that has actually been done)

Identity

- Sometimes, anonymous data can be linked to a specific person
- Other times, behavior identifies you
- Linkages can be used to establish identity
- MetroCards are anonymous—but credit cards aren't

Authentication

- If you're an authenticated user, your behavior can be tracked more easily over time
- (This includes Google, many media sites, etc.)
- Sometimes, even free accounts ask for demographic information, to improve profiles and ad targeting

"On the Internet, Nobody Knows You're a Dog"

- (Famous *New Yorker* cartoon)
- Often, what matters is not *who* you are, but what you do
- Example: for targeted ads, your identity doesn't matter, your interests do

Online and Offline

- You're profiled online and in the physical world
- Sometimes, the two are linked
- Profiling isn't new—but people have gotten a lot better at it

Offline

- Credit reports
- Credit card purchases
- Loyalty card programs
- Magazine subscriptions
- Warranty cards
- Public data (e.g., mortgages)
- Zip code demographics

Online

- Cookies
- “Flash local storage”
- Third-party ad sites
- Wireless ISP tracking headers

What's a Cookie?

- "Small text file stored on your computer"
- Set by a site; sent back to it next time you visit
- True—but frequently used to track you
- Persistent identifier, retained across sessions
- Not necessarily linked to a particular person—but the same each time you come back
- Linked to particular sites; one site can't retrieve another site's cookies

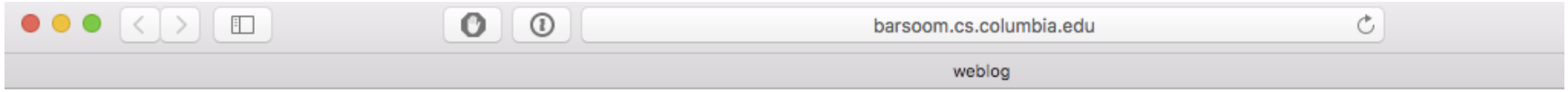
Good Uses for Cookies

- Login data
- Site preferences
- (Sometimes) shopping cart information

What Your Browser Reveals

- Ordinary: `http://greylock.cs.columbia.edu/`
- Advanced: `https://panopticlick.eff.org/`
- (Visit these sites on your own)

Safari Headers



I heard you say

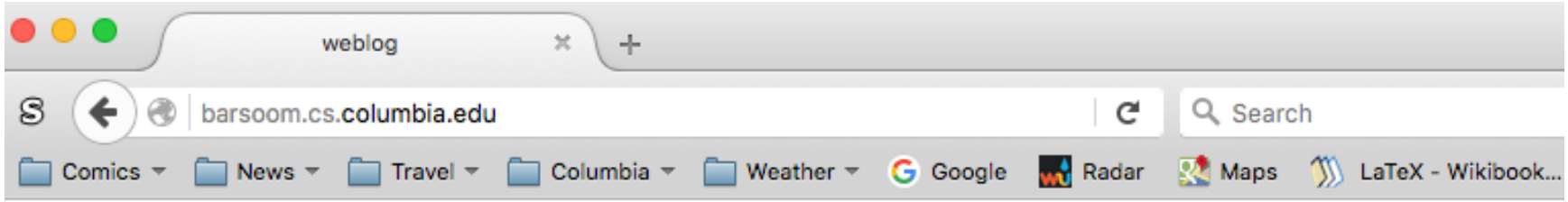
```
GET / HTTP/1.1
Host: barsoom.cs.columbia.edu
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-us
Connection: keep-alive
Accept-Encoding: gzip, deflate
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWebKit/601.4.4 (KHTML, like Gecko) Version/9.0.3 Safari/601.4.4
```

from 128.59.107.140:44074

I just sent you, #1889947178, a cookie; reload this page to see it coming back to me.

Chromebook Headers

Firefox Headers



I heard you say

```
GET / HTTP/1.1
Host: barsoom.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.11; rv:44.0) Gecko/20100101 Firefox/44.0
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
```

from 128.59.107.140:45223

I just sent you, #1477171087, a cookie; reload this page to see it coming back to me.

Tor Headers

I heard you say

```
GET / HTTP/1.1
Host: barsoom.cs.columbia.edu
User-Agent: Mozilla/5.0 (Windows NT 6.1; rv:38.0) Gecko/20100101 Firefox/38.0
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
Connection: keep-alive
```

from 94.242.246.23:54417

I just sent you, #1780695788, a cookie; reload this page to see it coming back to me.

Panopticlick

- Browsers leak lots of information
- Computers differ subtly from each other
- How much information is leaked?
- The EFF measured it with *Panopticlick*

Panopticlick: Safari

Your browser fingerprint **appears to be unique** among the 6,460,653 tested so far.

Currently, we estimate that your browser has a fingerprint that conveys **at least 22.62 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.

Panopticlick: Firefox

Within our dataset of several million visitors, only **one in 78787.2926829 browsers have the same fingerprint as yours.**

Currently, we estimate that your browser has a fingerprint that conveys **16.27 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**.

Panopticlick: Tor

Within our dataset of several million visitors, only **one in 6278.48104956 browsers have the same fingerprint as yours.**

Currently, we estimate that your browser has a fingerprint that conveys **12.62 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.

Third-Party Ad Sites

- Most ads on the web come from third parties, not the site you're visiting
- These third-party sites have their own cookies, which they set and receive
- If an ad site places content on multiple pages, they'll know which of those pages you visit; this lets them build up a very complete profile of your interests
- Sometimes, sites pass extra information about you to the ad providers
- One of the biggest ad providers is Doubleclick, which is owned by Google. . .

Federated Authentication

- Rather than requiring everyone to have a login on every site, use your Google or Facebook login to authenticate to other places
- Convenient—many fewer passwords to enter, remember, etc.
- But—Google, Facebook, etc., know what other sites you visit
- (Also security issues, but out of scope for this class)

Media Sites

- Many media sites, including at least the *New York Times* and the *Wall Street Journal*, track what types of articles you read
- This information is used for targeted advertising

Linking Online and Offline

- Online, it's easy to build a good profile of people
- If you buy something online, that site knows your name
- Use third-party cookies to associate your interest profile with a name

Credit Cards

- Most people have only a few credit cards
- If you use the same card for online and offline purchases, your physical person in a store can be linked to online behavior
- Special features have been put into some online payment protocols to facilitate this

Profiling: Good or Bad?

- Good: you see only ads you're interested in
- Bad: profiling is unpleasant. Besides, if you see interesting ads you're more likely to buy. . .

Fair Information Practices

- First “code of fair information practices” developed in 1973 at HEW
- 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

Safe Harbor

- The EU enshrines the FIPP into law (next class. . .), and bars export of data to countries that don't protect data well
- For the private sector, the US for the most part does not
- What about US companies doing business in Europe, but with data centers in the US?
- The old "Safe Harbor" provision let US companies store EU data if they promised compliance and if their promise was legally enforceable
- In the wake of the Snowden revelations, the ECJ invalidated Safe Harbor in October 2015
- The new Privacy Shield program has replaced it