Week 12 - Wednesday

COMP 4290

Last time

- What did we talk about last time?
- Exam 2!
- Before that:
  - Review
- Before that:
  - Started privacy

Questions?

Project 3

Aidan Kent Presents

Privacy Principles and Policies

Authentication

Authentication

- We have already discussed authentication from the perspective of how to do it
  - But what are we really authenticating?
- We could be authenticating any of the following three things:
  - **Individual**
    - The physical person
    - Example: you
  - **Identity**
    - A string or numerical descriptor
    - Examples: the name "Clarence", the account admin
  - **Attribute**
    - A characteristic
    - Examples: being 21, having top secret clearance

Individual authentication

- Authenticating an individual is difficult
- It mostly tracks back to a birth certificate
  - Which can be faked
  - Which contains very few characteristics that will not change over the years
- Some people fail to authenticate *themselves*
  - Believing they are someone else for years
- You acquire additional IDs and friends who can vouch for you
  - It's all a house of cards
- Perhaps DNA evidence can provide better individual authentication
- Plot of the classic Wilkie Collins novel (and many other thrillers) *The Woman in White*

Identity authentication

- Most authentication only authenticates an identity
  - If you have a credit card with a matching signature, you can buy stuff
  - If you have a student ID, you can swipe into the cafeteria
  - If you have a toll pass, you can drive through a toll
- When could each of these authentications give a false positive or a false negative?
- The biggest privacy danger is when outsiders can link relatively anonymous identities together

Anonymizing records

- The linkages between records can be the most dangerous
- Rich records are important for doing research
- As we discussed in the database chapter, it is very difficult to know how much data you can safely report
- Even "fully" anonymized records can leak who you are
  - Sweeney reports that 87% of the population of the US can be identified by zip code, gender, and date of birth
  - If medical research records include zip code, you can get pinned down
- What are the consequences?

Data Mining Privacy

Correlation in data mining

- Correlation is joining databases on common fields
- Privacy for correlation can be improved by making it harder to find links between related fields
- **Data perturbation** randomly swaps fields in records
  - Swapping records indiscriminately can destroy the value of the research
  - It has to add just enough randomness to the right fields

Aggregation in data mining

- Aggregation means reporting sums, medians, counts or other statistical measures
- As we discussed in the database chapter, these can threaten privacy if we have a very small sample size
  - If *n* items make up more than *k* percent of the subset, the subset is small enough to leak information
- A corresponding problem happens if we have a sample that includes almost but not quite all of the data
- For aggregates, data perturbation means adding small, random positive and negative values to each value, adding noise to the final aggregates
  - If done correctly, the aggregates may still be accurate enough for research purposes

Target example

- Like many big box stores, Target does extensive data mining on its customers
- Babies are big business, so Target sends ads for baby-related items to women it thinks are in their second trimester
- A father was angry and complained to Target that his daughter wasn't pregnant
  - And then found out that she was!

Privacy on the Web

The Internet

- As with everything else security-related, the Internet is why we have this course
  - The scale of computer security problems was simply much more manageable before the Internet
- The Internet allows some kinds of anonymity
  - But technology can allow you to be tracked in many ways
- Authentication (both of user and of service) can be difficult
- It is unclear what laws (if any) apply when you have your money or your identity stolen
- It is difficult to prosecute the criminals even if laws do apply

Credit card payments on the web

- Since cash isn't an option, many transactions use a credit (or debit) card
- To prevent fraud when accepting a credit card, merchants ask for:
  - Credit card number (of course)
  - Card security code (often called CVV)
  - Name on the card
  - Expiration date
  - Billing address
- Unfortunately, these pieces of data are exactly what would be needed for them to use your card
  - And this data can be used to correlate identities
- Many banks provided temporary credit card numbers that can only be used a fixed number of times or during a certain time frame
  - To be really careful, you probably should only use these

Payment schemes

- An alternative to credit cards are payment schemes like PayPal, Venmo (owned by PayPal), Cash App, and others
- The good:
  - You have more anonymity since the merchant only knows your e-mail address
  - Tight PayPal integration with eBay and other markets means you can often get your money back if an item never shows up
- The bad:
  - Consumer protection laws in the US usually limit your fraud liability on credit and debit cards to $50
    - PayPal does not have this kind of protection
  - PayPal still knows everything about you
- There are cell phone payment systems with similar issues
  - In the developing world, these payment systems are changing lives

Site registration

- Virtually every site on the Internet allows (if not requires) you to register with a user name and password so that you can log in
- For the sake of privacy, you should have a different ID and password for every site
  - This, of course, is impossible
- People tend to use one or two IDs (and one or two passwords) for everything
  - Many websites encourage this behavior by forcing you to use your e-mail address as your ID
- In this way, it is easy for anyone with access to multiple databases to aggregate information about you
  - Since your e-mail address is often tied closely to you, they could find out your true identity

Source of page content

- The book claims that site registration is in large part so that websites can give more detailed information to advertisers about who visits their sites
- As all of you know, websites are often filled with third party ads
- When you click on an ad (or use a coupon code listed on a particular ad), the advertiser knows what site you came from
  - This is how so much on the Internet is free, since targeted ads with feedback are better than TV or billboard ads
- Advertisers can learn a lot about the sites you visit and the products you buy

Cookies

- A **cookie** is a small text file kept on your computer that records data related to web browsing
  - It was originally intended to avoid the need to log on and store information on websites' servers
- Sites can store as many cookies as they want with any data they want (user name and password, credit card numbers, etc.)
- Cookies can only be read by the site that originally stored the cookie
- The way to get around this is called **third-party cookies**
  - Networks of sites can form an alliance in which they cooperate to track all of your visits to sites in the network
  - DoubleClick is such a network
- Tracking where you go online is called **online profiling**

Shopping on the Internet

- There are some good deals on the Internet
  - But there are also shady practices
- A typical brick-and-mortar company like McDonald's will sell everyone who comes into the store a cheeseburger for the same price
- Online stores may change prices on the fly based on previous browsing or buying histories
- Amazon.com had a differential pricing scandal when it was shown that loyal customers paid **more** in some cases
  - They have vowed not to do that anymore

Rights online

- Let's see how well we know the rules

Behavior	True or False	% Right on Survey
Most online merchants give you the right to correct incorrect information about you		
Most online merchants give you the chance to erase information about you		
It is legal for an online merchant to charge different people different prices at the same time of day		
It is legal for a supermarket to sell buying habit data		
Travel services such as Orbitz and Expedia have to present the lowest price found		
A video store is allowed to sell information about what videos a customer has rented (what a 1990s prompt!)		

E-mail Privacy

Interception of E-mail

- Regular mail cannot be opened under penalty of federal law
- Most people do not encrypt their e-mail using PGP or S/MIME
- Typical e-mail transmission:
  - Alice composes an e-mail on her computer
  - When she hits send, it goes to her organization's SMTP server
    - The organization can (and often does) keep a copy or at least scan the e-mail for questionable content
  - The SMTP sends it out through their ISP
    - Anyone on the Internet has a chance at grabbing the e-mail
  - It arrives at Bob's POP server
    - Bob's organization can record or scan the e-mail
  - Bob's computer pulls it down from the POP server and reads it

Monitoring of E-mail

- Companies and government agencies can legitimately monitor e-mail going to and from their employees
- The same is true for students at schools and patrons at libraries
- ISPs can monitor all the traffic that goes through them
  - They have to! So much e-mail in the world is spam
- You have no expectation of privacy when sending e-mail, ever

E-mail anonymity

- Some strategies can be adopted to maintain anonymity:
  - Sign up for a Gmail, Yahoo, or Hotmail account specifically to send a sensitive message
  - **Remailers** are trusted third parties who resend your e-mail under a pseudonym
    - But the remailer still knows who sent the e-mail
  - A **mixmaster remailer** takes it a step further by anonymizing through many layers
    - Only the first layer knows the sender
    - Only the last layer knows the receiver

E-mail authenticity

- Unless you verify authenticity cryptographically or through some other mechanism, you can't be sure where an e-mail comes from
- An e-mail is a series of packets, whose source IP address and from e-mail address can be spoofed
- Viruses also can take control of a computer and send e-mails to everyone on an address list
  - Sometimes they spoof the sender as someone else on the address list so that the virus is harder to track down

Upcoming

Next time…

- Finish privacy
- Security planning
- Abiral Pokharel presents

Reminders

- Finish Assignment 4
  - Due Friday!
- Work on Project 3
  - Phase 1 due next Friday
- Read Chapter 10