Week 11 - Wednesday

COMP 4290

Last time

- What did we talk about last time?
- Data mining
- Cloud computing

Questions?

Project 3

Assignment 4

Ashley Gutierrez Presents

Privacy Concepts

What is information privacy?

- Controlled disclosure
  - Right to control who knows your private data
  - Control is always diminished by sharing data with another party
- Sensitive data
  - Not all data is equally sensitive
  - Different people in different circumstances may disagree about what should be protected
- Affected subject
  - Both people and businesses have private data
  - Increasing privacy (an aspect of confidentiality) often decreases availability

Kinds of personally sensitive data

- Everyone has different standards of what is and isn't sensitive
- Don't forget that not everyone shares the same standards as you
- Potentially private data:
  - Identity
  - Financial and banking details
  - Legal issues
  - Medical conditions, drug use, DNA
  - Voting and political activism
  - Religious and sexual preferences
  - Biometric data
  - Diaries, poems, and journals
  - Confidential discussions with lawyers, accountants, doctors, and clergy
  - School or employment performance
  - Reading, web browsing, music listening, television and movies
  - Travel data and past citizenship
  - Mail, e-mail, SMS, and phone communication
  - Actions and history from the past, "youthful indiscretions"
  - Illegal activities, criminal records

Computer-related privacy problems

- Data collection
  - Computers make it possible to collect and store vast quantities of information
- No informed consent
  - Much data is now collected by websites or phone companies without announcement
  - Visiting a single website can be logged by your computer, the ISP, and the website itself
- Loss of control
  - Sharing information = lost control
  - With past forms of communication, there was more time to think and reflect about your message
  - It was easier to destroy the evidence, too
  - Data on the web (including Facebook) may never truly disappear
- Ownership of the data
  - Facebook (and others) own the data you post on their sites
  - They make money from your data

8 dimensions of privacy

- Rezgui et al. came up with 8 dimensions of privacy that relate specifically to computer use
  1. Information collection
     - Data is collected only with knowledge and consent
  2. Information usage
     - Data are used only for certain specified purposes
  3. Information retention
     - Data are retained only for a set period of time
  4. Information disclosure
     - Data are disclosed only to an authorized set of people
  5. Information security
     - Appropriate mechanisms are used to ensure the protection of the data
  6. Access control
     - All modes of access to all collected data are controlled
  7. Monitoring
     - Logs are maintained showing all accesses
  8. Policy changes
     - Less restrictive policies are never applied after the fact to preexisting data

Privacy Principles and Policies

Fair information policies

- In 1973, a committee advising the U.S. Department of Human Services proposed a set of principles for fair information practice:
  - Collection limitation
    - Only get data legally
  - Data quality
    - Data should be relevant, correct, and complete
  - Purpose specification
    - The purpose for the data should be identified and the data destroyed if no longer needed for that purpose
  - Use limitation
    - Use for other purposes only with the consent of the subject (or by authority of law)
  - Security safeguards
    - Procedures to protect the data should be in place
  - Openness
    - How the data systems work should be publicly available knowledge
  - Individual participation
    - The data subject should normally have a right to access and change his or her data
  - Accountability
    - A data controller should be accountable for making sure these principles are met

Data storage problems

- Realizing that large quantities of data storage would represent a threat, Ware and Turn suggest four ways to protect it:
  - Reduce exposure by limiting the data stored
    - Use random samples instead of complete surveys
  - Reduce data sensitivity by interchanging data items or adding small errors
  - Anonymize the data by removing or modifying identifying data
  - Encrypt the data

U.S. privacy laws

- The 1974 Privacy Act is a broad law that covers all the data collected by the government
  - The law is based on the principles from two slides earlier
- Laws for data collected by other organizations are for specific areas and not necessarily consistent
  - Fair Credit Reporting Act is for consumer credit
  - Health Insurance Portability and Accountability Act (HIPAA) is for healthcare information
  - Gramm-Leach-Bliley Act (GLBA) is for financial services
  - Children's Online Privacy Protection Act (COPPA) is for children's web access

Government websites

- In 2000 the Federal Trade Commission (FTC) mandated that government websites would have to address five privacy issues
  - Notice
    - Information policies must be disclosed before collecting data
  - Choice
    - Consumers must be given a choice about whether and how their information can be used
  - Access
    - Consumers should be able to view and contest their data
  - Security
    - Reasonable steps should be taken to protect the data
  - Enforcement
    - A mechanism for punishing noncompliance should be put in place

e-Government Act of 2002

- The e-Government Act of 2002 requires federal government agencies to post privacy policies on their websites
- The polices must disclose:
  - The information that is to be collected
  - The reason the information is being collected
  - The intended use by the agency of the information
  - The entities with whom the information will be shared
  - The opportunities for consent that would be provided to individuals regarding what information is collected and how that information is shared
  - The way in which the information will be secured
  - The rights of the individuals under the Privacy Act and other relevant laws

Commercial websites

- These standards do not apply to commercial websites
  - Specific industries are covered by the Fair Credit Reporting Act, HIPAA, GLBA, and COPPA, but most websites are not
- The FTC can prosecute companies for deceptive practices
  - Companies have to abide by their privacy policies
  - If a company does not list a privacy policy, they can do anything
  - If the privacy policy says they can sell your data to your worst enemy, they can

Examples of deceptive practices

- CartManager
  - Wrote software to manage shopping carts
  - Collected data from users even on commercial websites which claimed not to
  - Since the use of CartManager code was transparent to the user, they broke the law
- Jet Blue
  - Said it would not disclose passenger data to third parties
  - It gave credit card data to Torch Concepts which gave it to the Department of Defense
  - The DoD analyzed the data, looking for potential terrorists
  - Jet Blue violated its own policy, and the DoD avoided the government's own laws by collecting from a company instead of from individuals

Non-US privacy

- The European Union adopted the European Privacy Directive that requires that data about individuals be:
  - Processed fairly and lawfully
  - Collected for specified, explicit, and legitimate purposes
  - Adequate, relevant, and not excessive for the purposes they were collected
  - Accurate and as up to date as necessary
  - Kept in a form that permits identification of individuals for no longer than necessary

More European privacy

- The EU has three more principles for privacy:
  - Special protection for sensitive data
    - Including race, ethnicity, political opinions, religious beliefs, and health or sexual life
  - Data transfer
    - Authorized users of data cannot transfer it to others without permission of the subject
  - Independent oversight
    - Entities that process data (including the government) should be subject to independent oversight
- Since these principles apply to governments and businesses, they are stronger than US laws
- This causes problems since the EU laws make it illegal to share data with companies or governments (like us) with weaker privacy laws

Anonymity

- One way to avoid giving up private data is to conduct as much business as possible anonymously
- The Internet allows people to do a lot more anonymously
  - Some speculate that Amazon only survived because it created a private way to buy books about sex and porn
- Trusted third parties are needed to carry out business anonymously

Multiple identities and pseudonymity

- We all have multiple identities
  - Different credit card numbers, login information at millions of different websites, numerous pieces of government ID
- Some of this data (e.g. your name) can be used as a database key
  - Names are misspelled
  - Some people have common names (which can also be the same as terrorists)
  - Names can be changed
- Linking identities together correctly creates privacy risks
- Linking them together incorrectly causes other problems
- **Pseudonymity** is using disposable identities (junk e-mail addresses, etc.)
  - Artificially creating multiple identities
  - Can protect some privacy

Government and privacy

- The government collects a lot of data
  - Sometimes it collects it from private companies
- It also plays an important role in authentication
  - Issuing passports and drivers' licenses
- Risks from collecting data from outside sources:
  - Data errors
  - Inaccurate linking of identities
  - Difference of form and content
  - Purposely wrong data
  - False positives
  - Mission creep
  - Poorly protected integrity

Protecting against privacy loss

- The same committee that listed the risks on the previous slide suggested the following ways that the government should safeguard data:
  - **Data minimization** – Get the least data needed
  - **Data anonymization** – Replace identifying information with untraceable codes
  - **Audit trail** – Record who accesses what
  - **Security and controlled access** – Protect the data
  - **Training**
  - **Quality** – Consider the reason a particular source collected the data, how it was stored, how old it is
  - **Restricted usage** – Review if proposed uses of the data are consistent with why the data was collected
  - **Data left in place** – Leave the data with the original owner if possible
  - **Policy** – Establish a clear policy
- These recommendations have not (yet) been made law

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?

Ticket Out the Door

Upcoming

Next time…

- Review for Exam 2
- Olivia Crespo presents

Reminders

- Work on Project 3
  - **Passwords and phrases due this Friday in class!**
- Work on Assignment 4
  - Due Friday of next week
- Study for Exam 2
  - Monday in class