DS 300 Midterm 1 Study Guide

Josepeh Sepich Sept 27

1 Privacy Overview

Data is generated everywhere.

- Every click on-line
- user-generated
- Internet of things
- Health and Scientific computing

A 2013 study researchers were ablet o derive all the demographics of a person from a simple image on their social media account.

In the adversarial model sometiems the data controllers "dehumanize the enemy by calling them data".

- To what extent can someone use data beyond the person's intentions?
- Who owns data?
- How do we value different aspects of privacy?
- How do users get informed consent?
- What does it mean to be fair?

2 Data Science and Data

Predpol is a predictive policing service. The company adaptive an earthquake prediction algorithm to help predict crime hot spots that need police attention. In LA there's been a 33% reduction in buglaries and 21% reduction in violent crimes in areas where the software is being used. Some details that are privacy conerning are:

- area specific data from crime incidents is used
- sensitive data could be derived from non-sensitive data

Macy and other companies have also devloped price chaning algorithms. These algorithms essentially change prices for products in real time adapting to demand and inventory.

Yet another application of data science is Google's sponsored search. Google makes a revenue of $50 billion a year and 97% of that revenue comes from marketing. The sponsored search service uses an auction. This creates pure competition where companies try to win access to customers. Using data given by Google companies creating auction algorithms to have the right model in order to reach the right audience. Google has 30 billion requests a month.

Other applications of Data Science:

- Transaction Databases: Reccommendation system, Fraud detection
- Wireless Sensor Data: Internet of Things
- Text and Social Media Data
- Software Log Data: Automated Trouble Shooting

- Genotype Data: 23andMe, personliazed medicine

Five V's of Big Data

1. Volume: Raw Data
2. Velocity: Change over time
3. Variety: Data types
4. Veracity: Data quality
5. Value: Information for decision making

It is hard to balance utility of data and privacy

Privacy: State of being let alone and able to keep certain matters to one's self.

Types of Privacy:

- Personal Privacy
- Information Privacy
- Legal Privacy

In a Ted Talk by a researcher on privacy, there was an experiment performed at CMU. In this experiment the reserachers took a picture of someone and had them fill out a survey. While they filled out a survey they scraped the web with the picture. They were able to identify 1 in 3 people and could find SSN based of a picture and social media profiles.

Case Study: OkCupid Data

- Who owns the data?
- How do we value different aspects of privacy?
- How do we get informed consent?
- What does it mean to be fair?

Consider:

- Consent
- Transparency
- Terms of Service
- Adversarial Model
- Limits on inference: Where do we draw the line?

3 Overview of Data Privacy Issues

Privacy is a hard to define concept. Some examples of what it could be:

- Personhood
- Intimacy
- Secrecy
- Right to be let alone
- limited access to the self
- control over information

In this class we are focusing on privacy and not security, but security is a part of keeping data private.

Some inherently identifying information cannot be anonymized:

- genome sequences
- ancestry.com

Desires about data involve questions of trust. People must be willing to share information such as medical records, but are more likely to share with researchers and not pharama companies. Privacy is **not a binary value**.

Limited Acces:

- Laws to prohibit or limit collection, disclosure, contact
- tech to facilitate anonymous transactions

Control:

- Laws to mandate choice
- facilitate informed consent
- enforce privacy preferences

There are four facets of privacy:

1. Solitude: Separate from group, free from observation
2. Intimacy: put in a small unit
3. Anonymity: public but finds freedom from id/surveillance

Information privacy deals with collection of data and how it is stored. Decisional privacy deals with how end users decide to release or keep data.

Privacy and stakeholders:

- Company
- Customer
- Government
- data analyst
- data snooper/adversary

Confidential data definitions:

- Easily connected to a person providing it
- Could be agree to be kept confidential
  - business income
  - health/medical details
- conditioned by a number of factors
  - ethical guidelines
  - legal requirements
  - research specific consent agreement

3.1 Data Analysis Workflow

1. Collection of Data
2. Filtering and Processing of Data
3. Analysis of Data
4. Publications and usage of data

3.2 Privacy Violations

1. Information leakage (cookies on the web)
2. Inference (predictive analysis)
3. Information/data used in unwanted way (reccommendataions too personal)

Privacy Enhancing Technology

- VPN
- DuckDuckGo
- Wickr

Privacy Invasive Technology

- Amazon Echo
- Facial Recognition
- Location-based technology
- Venmo

4 Economics of Privacy

Sumamry

- Data vs Decisional Privacy
- Collection
- Accessor Usage
- Dissemination of Data

Key Questions

- Adversarial Model: Who is the adversary? What is their space of actions?
- Mechanisms: Are the right mechanisms in place to achieve the privacy goal?
- Incentives: Will human and economic factors favor or disfavor the privacy goal?

Economics of Privacy

- Cost and benefits associated with protection or disclosure of data
- Data subject
- Data holder
- society as a whole

On a societal level the cost of privacy is not realized until the data is disclosed. On an individual level however the individual has some privacy control on their personal sphere. Even though individuals are in control of their data, there is inconsistent behavior. People **want** privacy, but often are not willing to **take steps** to protect it. Some possible reasons are they do not care, the cost of protection is too high, they don't understand the implications of their behavior, or they want immediate gratification and the cost is not a tangible concept.

Sample experiment using gift cards illustrated the **endowment effect**. This effect states that if a person perceives to already posses something they want to retain what they already have.

Does control enhance or reduce privacy? It is important that individuals have control, but think of the airplane versus the car. When you are in a car you are in control and feel safer, but statistically are in much more danger. This could be the same when it comes to privacy controls.

Key takeways are that people often don't understand the implications of their actions. When making cost benefit analysis people can only see the tangible consequences in front of them and often do not understand the full spectrum of outcomes that can come from their actions. This is also a reason why sometimes people are perceived to act irrationally.

4.1 Privacy Laws around the World

Laws and regulations around the world vary. The US has a "patchwork quilt" when it comes to privacy laws and regulations. Laws and regulations are sector specific and often have minimal protections. Some regulators in the US are teh Federal Trade Commission and the Federal Communications Commission.

4.2 Fair Information Practice Principles (FIPS)

1. Collection Limitation
2. Data Quality
3. Purpose Specification
4. Use limitation
5. Security Safeguards
6. Openness
7. Individual Participation
8. Accountability

5 Laws and Directives

5.1 European Data Protection

In Europe each country has their own commissions, but there is also a single law governing the entire union created in 2018: GDPR, General Data and Privacy Regulations.

5.1.1 GDPR

- Doesn't matter where data is processed
  - Origin of data is Europe
- Explicit Consent must be given
- Data management
- Retained control
- Care for sensitive data
- Similar pillars to FIPS

5.2 FIPS

5.2.1 Collection Limitation

Collection limitation states that data must be collected for a specific purpose, so don't collect data not useful for that purpose.

5.2.2 Data Quality

The data that is stored should stay consistent and accurate.

5.2.3 Purpose Specification

Data is collection for a specific purpose.

5.2.4 Use Limitation

The data is only used for the intended purpose.

5.2.5 Security Safeguards

The proper security protections should be put in place to prevent data leaks.

5.2.6 Openness

Data holders must be open about their policies.

5.2.7 Individual Participation

Users should help contribute to data management.

5.2.8 Accountability

The data holder is responsible for their actions.

5.3 Privacy Risks in E-Commerce

- Unsolicited Marketing
- Subpoena
- Government Surveillance

One should start privacy analysis by identifying risks. What are the consequences and who could obtain the data?

6 Protecting Sensitive Data