Data Security and Data Privacy

Miranda Braselton - Carlos Rojo - Carol Stanton

Data Security and Data Privacy

What are they and how do we teach students about these concepts?

Overview

- Differences between data security and data privacy
- Introduction to data privacy
- Introduction to cybersecurity
  -malware detection as an example
- Intro to teaching students about data privacy

Data security vs. Data Privacy

Data Security

↓

Prevent unauthorized access to *private* databases or software

↓

E.g. hacking, spam email, malware prevention

Data Privacy

↓

Prevent identification of individuals in *released* databases

↓

E.g. remove name, zip code, social security number from data records

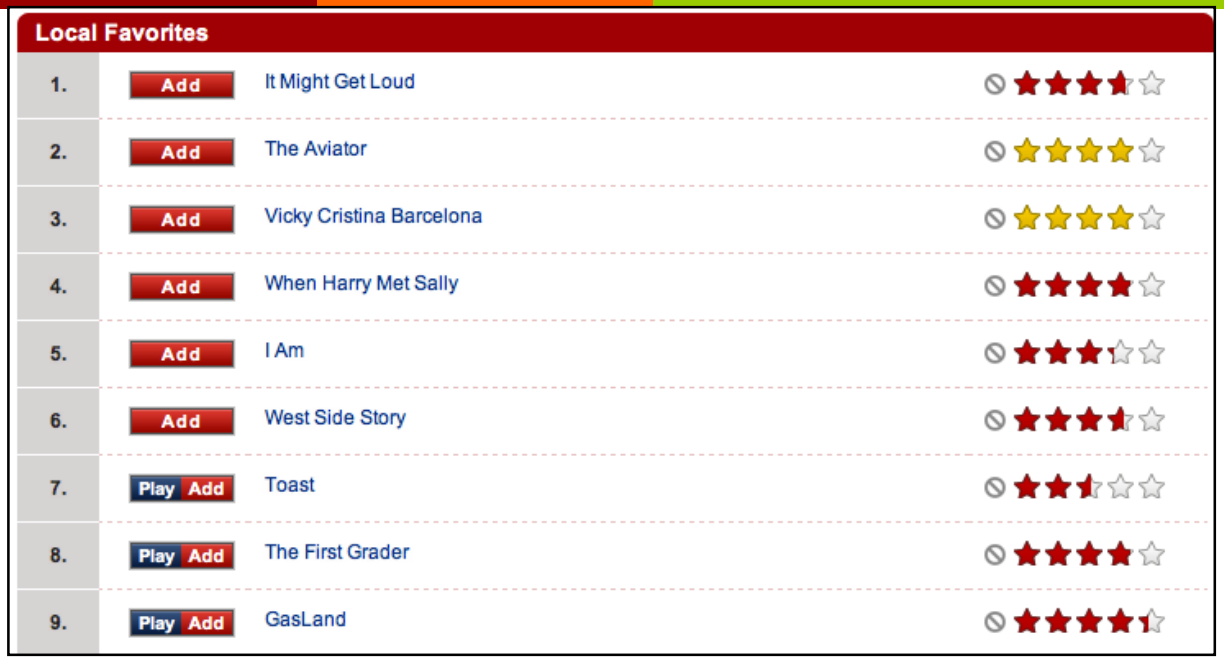
Data Privacy

- Many publicly available datasets contain potentially sensitive information (income, disease status, etc.).

- Data privacy is concerned with ensuring that when we release a dataset, the individual that owns the data cannot be uniquely identified beyond a certain threshold.

- Data privacy or anonymity is often attempted by "de-identifying" the data (i.e. removing names, social security numbers, zip codes from data records).

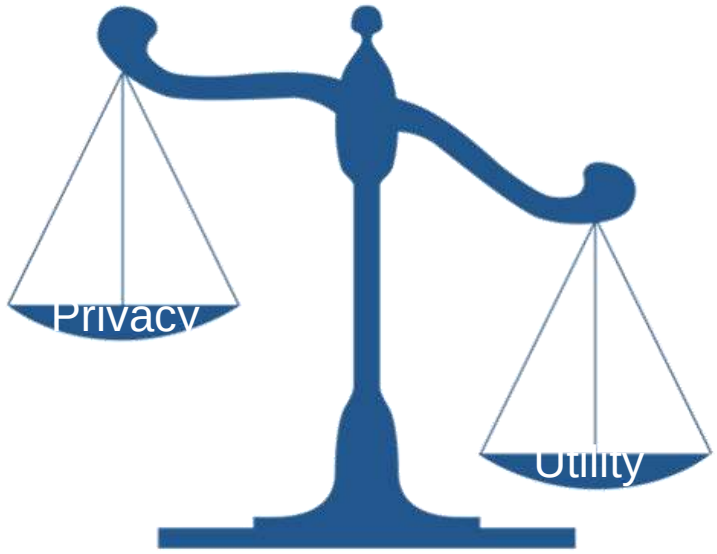
De-identification often not enough



Netflix releases de-identified (and supposedly privatized) user ratings dataset. Researchers are able to use this public dataset to assign the ratings to unique individuals and determine what movies they were watching (a serious breach of privacy)

http://www.wired.com/2009/12/netflix-privacy-lawsuit/

Strategies to better privatize data



- k-anonymity

-ensure that at least k records in the dataset have the same values (preventing unique identification of any individual)

- Do not provide exact values when releasing dataset

-E.g. if releasing ages, provide ranges. Instead of "36", release age as "35-40"

- Note that every privatization strategy increases privacy BUT *decreases* data utility since privatization removes information from the dataset.

Overview

- Differences between data security and data privacy
- Introduction to data privacy
- Introduction to cybersecurity
  -malware detection as an example
- Intro to teaching students about data privacy

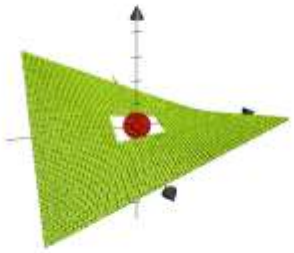
Cyber Security

Anomaly Detection

- The ability to identify anomalous data, quickly and accurately, is of critical importance in many fields.
- Medicine
    - Testing and diagnosis
- Environmental science
    - Air and water quality
- Engineering
    - Structural failure
- Finance
    - Fraud
- In the area of cyber security, anomalies can represent
- Malware
- Network intrusion

Anomalous Data

- Data
  - A data instance consists of a number of nominal and numeric attributes.
  - A dataset is a collection of data instances.
- Anomalies in a dataset
  - Data instances that are few in number and different from normal instances.
- Illustration
  - The points in the sphere are few and from the points on the saddle.



nt

Anomaly Detection Methods

- Statistical
  - Univariate statistics are computed for each attribute to create a profile for normal data.
  - Anomalies are significantly different from profile.
- Distance/density measures
  - Distance to nearest neighbor(s) is computed.
  - Anomalies are far away from nearest neighbor.
- Trees
  - Data instances are separated via binary decision trees.
  - Anomalies tend to be isolated higher in the trees.

Overview

- Differences between data security and data privacy
- Introduction to data privacy
- Introduction to cybersecurity
  -malware detection as an example
- Intro to teaching students about data privacy

Teaching Privacy

- The growing use of social-networking sites along with technological advances in data-retrieval techniques are providing new opportunities to make use of people's personal information — both ethically and unethically.

- It is becoming more and more important to provide education to individuals at an early age, about the many aspects and issues regarding online privacy.

TROPE:
Teachers' Resources for Online Privacy Education

"This project aims to empower K-12 students and college undergrads in making informed choices about privacy, by building a set of educational tools and hands-on exercises to help teachers demonstrate what happens to personal information on the Internet — and what the effects of sharing information can be."

Berkeley
UNIVERSITY OF CALIFORNIA

www.teachingprivacy.org

ICSI
INTERNATIONAL COMPUTER SCIENCE INSTITUTE

TROPE Online Privacy Topics

- **You're Leaving Footprints:** Your information footprint is larger than you might think.
- **There's No Anonymity**: Even just a part of your information may make it possible for someone to uniquely identify you even without your name.
- **Your Information Is Valuable:** Every piece of information, public or not, has value to somebody, and they will use that information however benefits them.
- **Sharing Releases Control**: When any information is shared online, you can't control what happens to it — or how people will interpret it.