CS 594: Empirical Analysis

Deriving Sound Insights from Data

Session 4: Privacy concerns

Today's class

- Discuss reading
- Homework 1 initial feedback
  - What meaningful questions (even if trivial) could we ask of this data?
- Dataset brainstorming
  - What interesting questions could be asked of dataset X?
  - How can we validate that we're able to answer that question?

Robust De-anonymization of Large Sparse Datasets

Narayanan and Shmatikov

Auxiliary database:

A subset of additional data that contains identifiers.

The netflix prize dataset includes a subset of all users. What was netflix's claim about how subscribers were chosen for this dataset?

How did the authors attempt to verify that claim?

- What's the goal behind showing this graph?
- What hypotheses might you want to test about this claim?
- Can we design an experiment that would allow us to test this?

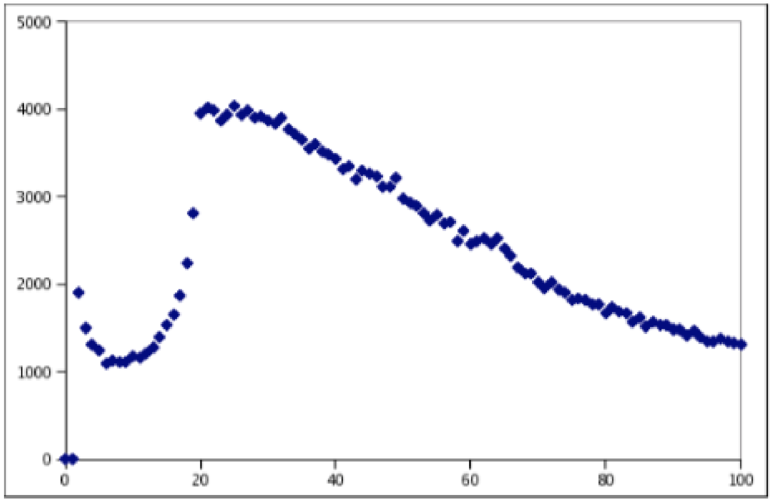


Figure 2. For each $X \leq 100$, the number of subscribers with X ratings in the released dataset.

The authors performed a large number of experiments, each of which needs the main database and the auxiliary database. What datasets did they use for this task?

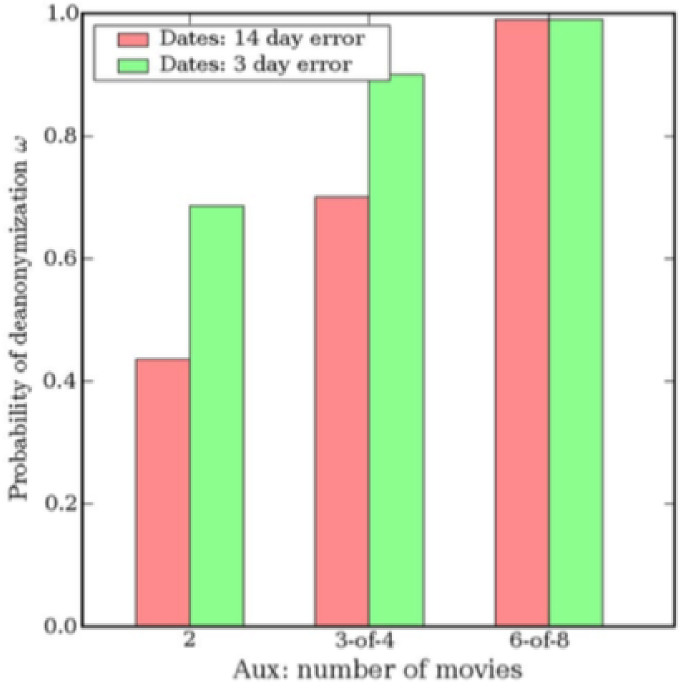
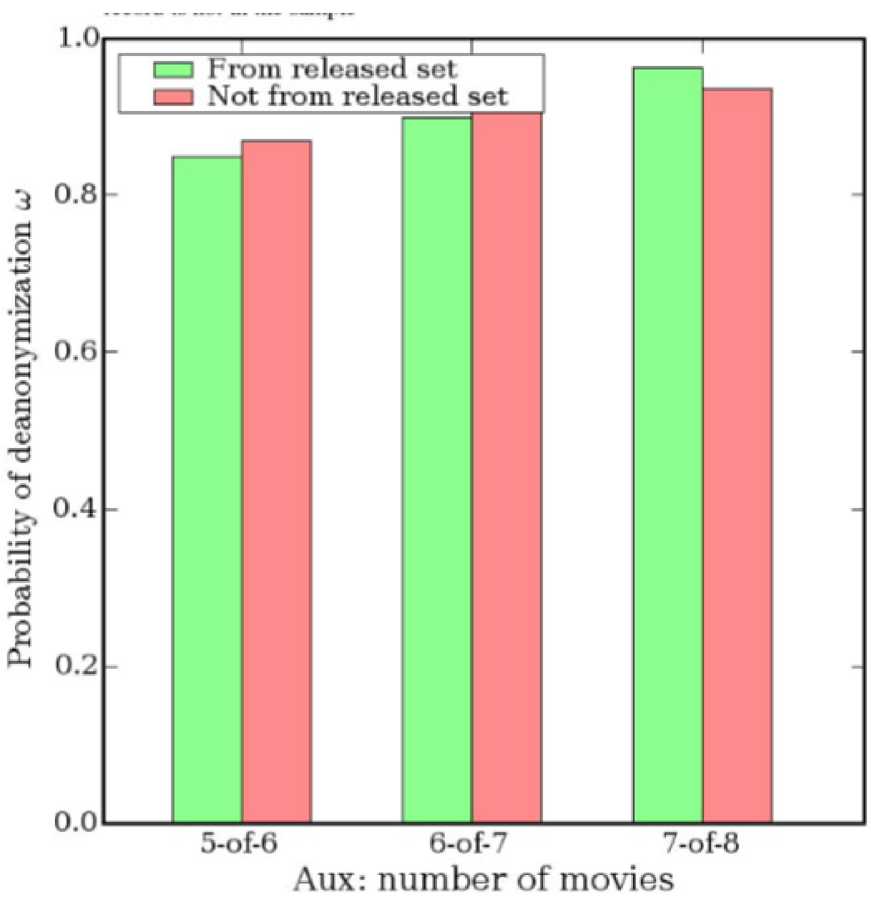


Figure 4. Adversary knows exact ratings and approximate dates.


1.0
0.8
0.6
0.4
0.2
0.0
Probability of deanonymization ω
🟩 From released set
🟥 Not from released set
5-of-6
6-of-7
7-of-8
Aux: number of movies

What variables did the experimenters vary during the course of their experiments?

What additional test datasets did the researchers do that used information beyond the netflix prize dataset?

- Dataset brainstorming
    - What interesting questions could be asked of dataset X?
    - How can we validate that we're able to answer that question?

- Enron email corpus
- CRAWDAD wireless networking repository datasets
- CAIDA internet measurements
- DNS Zonefiles
- Wikipedia
- StackOverflow
- City of Chicago open data
- FTC fraud complaint list
- Activitst data