Unlinkability: A Data Science Perspective

Ed Felten

Department of Computer Science
Woodrow Wilson School of Public and Intl. Affairs
Center for Information Technology Policy
Princeton University
Why we’re discussing this

Want definition of “unlinkable” data

Rationale: okay to override user’s DNT choice, if user has no real privacy interest at stake

Complements other exceptions
Available design space

data utility

privacy
Available design space

data utility

privacy
What does it mean for a data operation to be “privacy preserving?”
What does it mean for a data operation to be “privacy preserving?”

40+ years of research on this question

intuition is an unreliable guide
Intuition says:

If you’re not in the dataset, it can’t convey info about you.
Intuition says:
   If you’re not in the dataset, it can’t convey info about you.

Actually:
   Dataset can convey info about people who aren’t in it.
   That can be a good thing.
Intuition says:
  If you’re not in the dataset, it can’t convey info about you.

Actually:
  Dataset can convey info about people who aren’t in it.
  That can be a good thing.

Example:
  You are a smoker.
  Dataset shows that smoking increases cancer risk.
Intuition says: aggregate data is always safe

But:

Answer questions about yourself
Get recommendations
Aggregate data over >1M users

Researchers: can recover nearly all individual data

Similar demonstrations:
What does it mean for a data operation to be “privacy preserving?”
analyst

raw data

query

answer

query

answer

side information
Minimal requirements for a definition:

• Feasible
• Technically actionable
• Does not ban all data release
• Implies some limit on data inference

Modest goals, but very hard to achieve!
K-Anonymity

K-anonymity (aka “large bucket size”) fails to meet requirements

Why?

does not limit data inference
assumes only one query
assumes no side info
does not speak to all cases (e.g. aggregate data)
Dalenius’s Goal

What analyst learns about you, given side info + answers

≈

What analyst learns about you, given side info only
Dalenius’s Rule

What analyst learns about you, given side info + answers

≈

What analyst learns about you, given side info only

Fails – not feasible
Differential Privacy

Answer based on the full dataset

≈

Answer if your data wasn’t there
Differential Privacy: Advantages

- only definition known to meet requirements
- intuition for user: anything bad that happens would have happened anyway even without your data
- adjustable “leakage level” – a knob to turn, to trade off privacy vs. utility
- multiple queries: leakage combines additively
- not affected by side information (enhancement)
- known methods exist to achieve DP for (e.g.) aggregate counting queries