Big Data L12: Socio-cultural Impact

Topics Covered:

- Socio-cultural impact of recommender and information retrieval systems
- Privacy issues and de-anonymization
- Introduction to Differential Privacy

1 Socio-cultural Impact of Recommender Systems

Claim:
Recommender systems might have contributed to the downfall of sites like Buzzfeed.
Issues:

- **Echo Chambers and Filter Bubbles**:
  - Recommendations rely on **similarity**.
  - Over time, diversity decreases; users herd around similar content.

Pariser, 2011 coined "filter bubble."

    - Best case: users get what they like.
    - Typical case: users get bored and leave.
    - Worst case: users become **isolated** and **polarized**.

- **Targeted Advertising**:
  - Personalization often based on user features (Age, Gender, Zip code).
  - Can lead to discrimination issues (e.g., lawsuits against Facebook).
- **Ethical Challenges in Representation**:
  - Recommenders are shaped by **biased past behaviors**.
  - Important questions:
    - How is the model biased?
    - How are atypical users treated?
    - Who truly benefits from personalization?

2 Privacy and De-anonymization

Problems with Anonymization:

- Simply **removing identifiers** (names, IDs) isn't enough.
- **Common but flawed methods**:
  - **Obfuscate identifiers**: Replacing names with random numbers.
  - **Perturb observations**: Adding random noise.
  - **k-anonymity** [Sweeney, 2002]: each attribute shared by at least k people.
  - **Only publishing summary statistics**: Still leaky!

3 De-anonymization Attacks

Netflix Prize Attack ([Narayanan & Shmatikov, 2008]):

- Netflix released "anonymized" movie rating data.
- Attack:
  1. Define **similarity** between users.
  2. Given **partial ratings**, compute similarity to users.
  3. If the match is strong, re-identify the user.
- **Result**:
  - With just **8 ratings** (allowing 2 mistakes) and 14 days timestamp fuzziness, **99%** of users could be uniquely identified!
  - Even **without timestamps**, rare movie ratings leak identity.

Why it matters:

- Preferences (movies, music) correlate with sensitive personal attributes (politics, religion, sexual orientation).
- **Privacy breaches are irreversible.**

4 Broader Examples of Privacy Violations

- **2010 US Census Attack** ([Abowd, 2019]):
  - Reconstruction attacks using public census summaries.
- **Target Pregnancy Prediction** ([Duhigg, 2012]):
  - Target inferred a teenager's pregnancy based on shopping patterns, disclosed it accidentally.

5 Differential Privacy (DP)

Concept:
If one individual's data is removed, the result of any computation should not substantially change.

- **DP is a property of algorithms, not datasets.**
- **Randomization happens at the algorithm level**, not by modifying the raw data.

Formal Definition:
For datasets D and D' differing by one record:

$$\Pr[A(D) \in S] \leq e^{\epsilon} \times \Pr[A(D') \in S] \quad (1)$$

where:

- ϵ (epsilon) controls **privacy loss**.
  - Smaller ϵ: Stronger privacy.
  - Larger ϵ: Weaker privacy but more accuracy.

Mechanism: Adding Laplace Noise

- **Sensitivity** (Δf): Maximum change one row can cause in the output.

- **Laplace mechanism**: Add noise drawn from Laplace$(0, \Delta f/\epsilon)$.
- **Why Laplace noise?**
  - Heavy tails → better protection compared to Gaussian noise.

Trade-off:

Noise Level	**Privacy**	**Accuracy**
High Noise (small ϵ)	High	Low
Low Noise (large ϵ)	Low	High

- **Larger datasets** → **easier privacy** (sensitivity decreases).

Differential Privacy in Action:

- **Sum queries** (e.g., total clicks) are less sensitive than **Max queries** (e.g., maximum income).
- Privacy loss **accumulates** over multiple queries!

6 Summary

- **Simply de-identifying data is not enough** — high-dimensional data is easy to re-identify.
- **Differential Privacy** offers a mathematically sound way to **release data while protecting individuals**.
- **Laplace noise** carefully balances **privacy and reproducibility**.