8-1: Threat Models

Introduction

- We've talked some about privacy, noise
- This video: recommender threats more generally
  - Privacy
  - Robustness
- Primarily focus on malicious behavior, but has implications for benign problems (such as inconsistent ratings)

Core Question

- What does it mean for a recommender to be *secure*?
  - Or robust
  - Or protect privacy

Threat Model

- Protect *something* (important to the recommender or its users)
  - from *someone*
  - who has *goals*
  - and certain *capabilities*

Example: Influence Limiter

- Protect recommender accuracy and neutrality
- From malicious users
- Who want to push or kill products
- And can create fake accounts

Influence Limiter Solution

- Require users to prove themselves; malicious users have threshold to cross
  - Make the system resilient to the users
- Alternative approach: detect and remove

Protect System Accuracy

- Protect recommender accuracy
- From users
- Who want to disrupt its quality (or just give low-quality, inconsistent ratings)
  - This is all users
- And can create profiles and ratings
- Normal de-noising problem (malicious or natural noise, they both fit in this framing)

Example: User-User Privacy

- Protect user data
- From other users of the system
- Who want to know users' opinions
- And can create profiles, manipulate ratings
- Attack: use Pearson correlation problems to identify users, get their ratings
- Mitigation: use less transparent algorithm

Example: User-System Privacy

- Protect info about user
- From the service provider
- Who wants to know user characteristics
- And can analyze all users' data
- This is hard!

User-System Privacy Ideas

- Separate recommender from vendor
- Use Trusted Computing to attest recommender integrity
- Pool ratings between users
- Add noise to ratings & profiles
- Decentralize recommendation
- Homomorphic encryption

Conclusion

- Think carefully about the threats you want to protect from
- Think about what threats your users might consider
- Define threat model carefully when making privacy claims

8-1: Threat Models