Membership Inference Attacks Against Recommender Systems (acm.org)

Abstract

- **Objective**: The paper aims to quantify the privacy leakage of recommender systems through membership inference attacks.
- **Approach**: The authors propose a novel method that differs from traditional membership inference attacks, focusing on **user-level inference** rather than sample-level. The attack is based on observing ordered recommended items from a recommender system.
- **Shadow Recommender**: They establish a '**shadow recommender**' to derive labeled training data for training the attack model.

2. Establishing a 'Shadow Recommender' and Deriving Labeled Training Data

- **Shadow Recommender Creation**: The shadow recommender is created to mimic the behavior of the target recommender system. It is trained on a shadow dataset which is separate from the target dataset.
- **Labeled Data Generation**: The shadow recommender generates labeled training data for the attack model. This data includes interactions (user features) and recommendations (recommendation features). The process involves vectorizing interaction and recommendation sets to create feature vectors for the corresponding items.

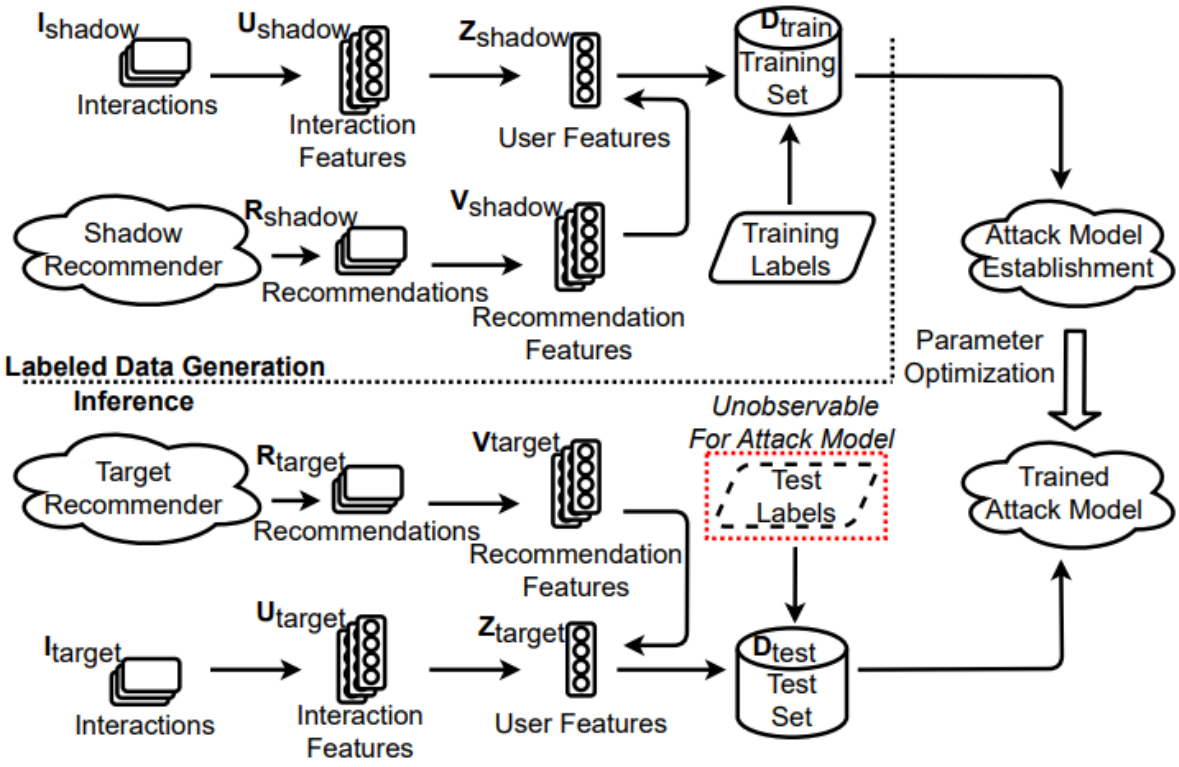

Figure 2: The framework of the membership inference attack against a recommender system.

- **Results**: Extensive experiments show that their attack framework is highly effective. They also design a defense mechanism to mitigate the threat.

Contributions

- **Privacy Risks Quantification**: The paper quantifies the privacy risks of recommender systems using membership inference, addressing two primary challenges: the attack is user-level, and the adversary only has access to the recommended items list, not the posterior probabilities.

1. Measuring Privacy Risks

  - **Metric Used**: The paper uses the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve as a metric to measure privacy risks. The AUC metric is commonly used in machine learning to evaluate the performance of binary classification systems. It provides a quantitative measure of how well

the attack model can distinguish between members (users whose data was used in training the recommender system) and non-members.

- **Attack Method**: The paper presents a new method to extract information from decision results for the attack model, considering that recommender systems output ranking lists of items.
- **Attack Execution**: The attack model, trained on the data generated by the shadow recommender, uses these representations to infer whether a particular user's data was used in training the target recommender system.

Framework of the Attack

- **Data Sets and Interactions**:
  - **Target Interactions (Itarget)**: This comprises a set of lists of user interactions, such as item ratings or purchase history, used by the target recommender system.
  - **Recommendations (Rtarget)**: This involves a set of ordered lists of recommendations generated for users by the recommender system.
  - The attack model is represented as $A: I\text{target} \rightarrow R\text{target}$, where $I\text{target}$ is the input (user interactions) and $R\text{target}$ is the output (recommendations).

Attack Process

1. **Data Collection and Preparation**:
   - The adversary collects interaction data and the corresponding recommendations for a set of users.
   - This data is used to create feature vectors that represent the users' interactions (Itarget) and the recommendations they receive (Rtarget).
2. **Building the Attack Model**:
   - The attack model aims to distinguish between members (users whose data was used to train the recommender system) and non-members (users whose data was not used).
   - To train the attack model, the adversary uses the data generated from the shadow recommender system, which includes both members and non-

members.

3. **Executing the Attack**:
    - The trained attack model is then applied to the target recommender system. It uses the observed interactions and recommendations of users to infer their membership status.
    - This process involves analyzing the patterns in the recommendations and comparing them with the known patterns from the shadow recommender system.

-

- **Threat Model**: The adversary's goal is to infer if a user's data is used in the target recommender. The assumption is that the adversary has black-box access to the target recommender, observing only the recommended items and the user’s history.

Components of the Threat Model

1. **Target Recommender**: This is the recommender system that is being attacked. It is trained on a specific dataset, referred to as the "Target Dataset."
2. **Shadow Recommender**: A separate recommender system built by the adversary. It is trained on a different set of data, known as the "Shadow Dataset." The purpose of the Shadow Recommender is twofold: to infer the membership status of users in the target recommender and to generate training data for the attack model.
3. **Members vs. Non-Members**:
    - **Members**: Users whose data is used to train the recommender system.
    - **Non-Members**: Users whose data is not used in the training of the recommender system.
4. **Recommendation Algorithms**:
    - **Personalized Recommendation Algorithms**: These algorithms learn a member’s preferences based on their historical behavior, such as purchases or ratings (termed as "Interactions").

- **Non-Personalized Recommendation Algorithms**: These algorithms are based on predetermined rules, such as selecting the most popular or highest-rated items.

5. **Recommendations**: Items recommended to users, varying based on the method of recommendation (personalized or non-personalized).
6. **Feature Vectors**: Represent latent features indicating item attributes or user preferences.
7. **Attack Model**: This model is central to the attack process. It aims to infer whether a target user is a member of the target recommender system. The model is trained on data generated from the Shadow Recommender.

Functioning of the Threat Model

- The threat model operates by using the Shadow Recommender to simulate the Target Recommender's behavior and create a training dataset for the Attack Model.
- The Attack Model then uses this data to make inferences about whether specific users' data were used in the training of the Target Recommender, thus determining their membership status.
- This process essentially leverages the differences in recommendations made to members and non-members, using these variations as clues to infer membership.

Defense Mechanism

- **Popularity Randomization**: To mitigate privacy risks, the authors propose 'Popularity Randomization'. This method involves enlarging the set of popular items and randomly selecting a subset for recommendation to non-member users.
- **Effectiveness**: Experimental results demonstrate that this approach effectively reduces the attack's success rate, particularly against certain algorithms like NCF (Neural Collaborative Filtering).

In summary, this paper provides significant insights into the privacy vulnerabilities of recommender systems and introduces a new approach for membership inference attacks that are effective in revealing privacy risks. Additionally, it proposes a defense strategy to mitigate these risks, offering a comprehensive view of the challenges and solutions in the domain of recommender system privacy.