Research on Privacy Verification in Federated Recommender Systems Accepted for Presentation at RecSys 2025, the Leading International Conference on Recommender Systems

Toyota Motor Corporation
InfoTech
Kenji Shinoda, Takeyuki Sasai, Shintaro Fukushima

Toyota Motor Corporation has identified a new vulnerability in federated recommender systems. The research paper “Popularity-Bias Vulnerability: Semi-Supervised Label Inference Attack on Federated Recommender Systems” has been accepted to the Nineteenth ACM Conference on Recommender Systems (RecSys 2025), a leading international conference in the field. The paper was presented at the main conference in September 2025.

Background

To realize human-centered AI agents, recommendation technologies grounded in user understanding are crucial. In particular, Vertical Federated Learning (VFL) has attracted significant attention. VFL enables privacy-preserving model training without sharing raw data among organizations. In VFL, each organization exchanges information such as gradients secured with strong cryptographic techniques (e.g., homomorphic encryption) and updates its local model. This approach is expected to allow models to be trained using not only an organization’s own data but also data from other organizations while preserving privacy.

However, this technology is still maturing, and attacks have been reported in which a malicious participant in VFL can infer and leak data that should remain protected, such as labels. To address these risks, ongoing research aims to develop practical and secure algorithms. To proactively assess security risks in industrial deployments, we investigated whether applying VFL to recommender systems causes domain-specific vulnerabilities. We show that a malicious participant can exploit characteristic properties of recommender systems to perform a label-inference attack that leaks the labels.

Overview

In many VFL settings, participating organizations fall into two roles. An active party that holds the training labels and one or more passive parties that provide only features. Because the active party has access to labels, it maintains a label-informed predictive model, whereas passive parties lack any independent label-inference capability.

Prior work has reported attacks in which a passive party can infer and leak the active party’s labels. One line of work uses a small set of leaked labels to train a semi-supervised model from its local model and to enable it to infer. Another line of work, in tree-based VFL, analyzes the local model’s tree to recover each sample’s leaf-node assignment for inference. To advance VFL toward industrial deployment, it is crucial to go beyond existing studies and uncover domain-specific vulnerabilities. Accordingly, this work investigates such vulnerabilities in the context of recommender systems.

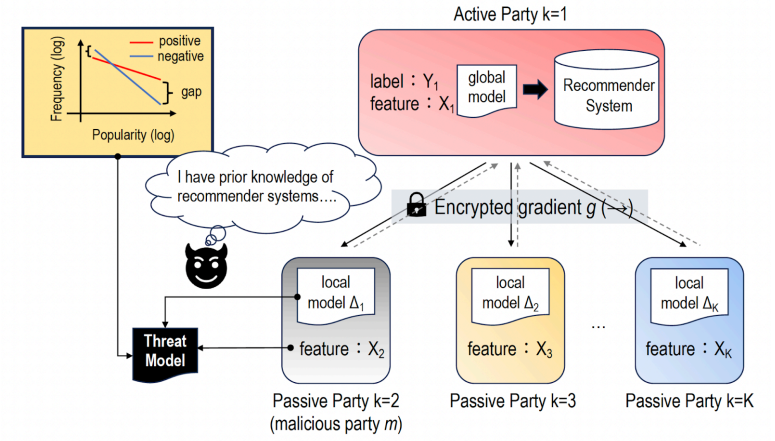

Illustration of a vertical federated recommender system with a malicious passive party.

In recommender systems, models are trained to predict user-item labels such as purchase or visit events. These logs reveal which items a user has clearly shown interest in. However, the absence of an interaction does not reliably imply disinterest in other items. This setting is referred to as implicit feedback. Supervised learning typically requires labels for both positive examples (items a user is interested in) and negative examples (items they are not). Under implicit-feedback settings, explicit negatives are unavailable and must be synthesized most commonly via random negative sampling.

Recommender systems often exhibit the following properties:

(1) Item popularity follows a power-law distribution: when plotting popularity on the x-axis and the number of items with that popularity on the y-axis, the distribution typically obeys a power law.
(2) Negative examples are generated by random sampling: under implicit feedback, it is common to sample, for each user, items the user has not purchased or visited and treat them as negatives.

When (2) is applied under (1), a pronounced skew arises between the distributions of positive and negative examples. Our key insight is that, if the malicious party exploits this skew as prior knowledge, they can exploit it to build the threat model.

We propose a three stage method targeting tree-based VFL with gradient-boosted decision trees such as SecureBoost.

Stage 1) Graph construction via tree analysis: Using the local model available to a malicious passive party, an edge is added between nodes (IDs, user–item pairs) if they are estimated to fall into the same leaf, thereby constructing a graph. The features held by the passive party are then attached to the nodes.
Stage 2) Leveraging prior knowledge: Using the skew between positive and negative distributions, the malicious party creates pseudo-labels and defines a prior probability for the loss.
Stage 3) Semi-supervised learning for the attack model: Using the graph from Stage 1 together with the pseudo-labels and prior from Stage 2, the malicious party performs graph-based semi-supervised learning to train the threat model.

This procedure enables a malicious passive party to infer the active party's labels. We validate the attack on three public datasets and confirm that it outperforms baseline methods.

Conclusion

This study reveals a new vulnerability in VFL-based recommender systems that exploits domain-specific properties, namely item popularity and the use of random negative sampling under implicit feedback. Proactively recognizing this attack will inform the design of effective defenses, helping to safeguard data and advance the development of human-centered AI agents.