

Aegis: Post-Training Attribute Unlearning in Federated Recommender Systems against Attribute Inference Attacks

Wenhan Wu¹, Jiawei Jiang^{1,*}, Chuang Hu^{2,*}
¹ School of School Science, Wuhan University, Wuhan, China

² State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, Macau SAR

*Corresponding Authors

Introduction and Motivation

Privacy Issues in Federated Recommender Systems.

Traditional recommender systems require centralizing user data, which poses privacy risks. Federated recommender systems (FedRecs) address this by keeping user data local, but user embeddings still risk exposing sensitive attributes, making them vulnerable to attribute inference attacks (AIAs). For example, a malicious client may be interested in all users' private attributes. This client could reach an agreement with the server and intentionally leak a portion of its users' private attributes, or this information might be inadvertently leaked.

Alternatively, a malicious server could collaborate with a curious client, where the client provides the server with private attribute data in exchange for some financial benefit.

Table 1: FedRec Recommendation Utility and Attribute Inference Attack Results on Different Datasets

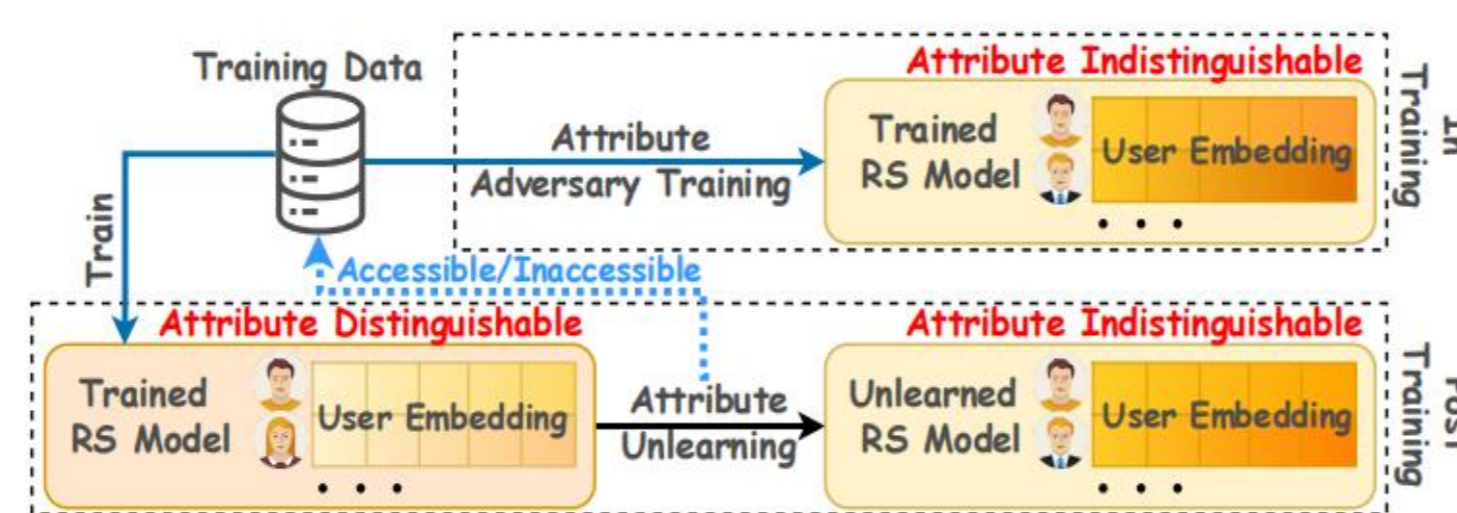
Dataset	Utility		Attribute Privacy		
	NDCG@10	HR@10	Gender	Age	Occupation
ML-100K	0.708	0.680	0.714	0.280	0.149
ML-1M	0.699	0.684	0.849	0.353	0.119
Random Attacker			0.500	0.143	0.048

Limitations of Existing Privacy Mechanisms.

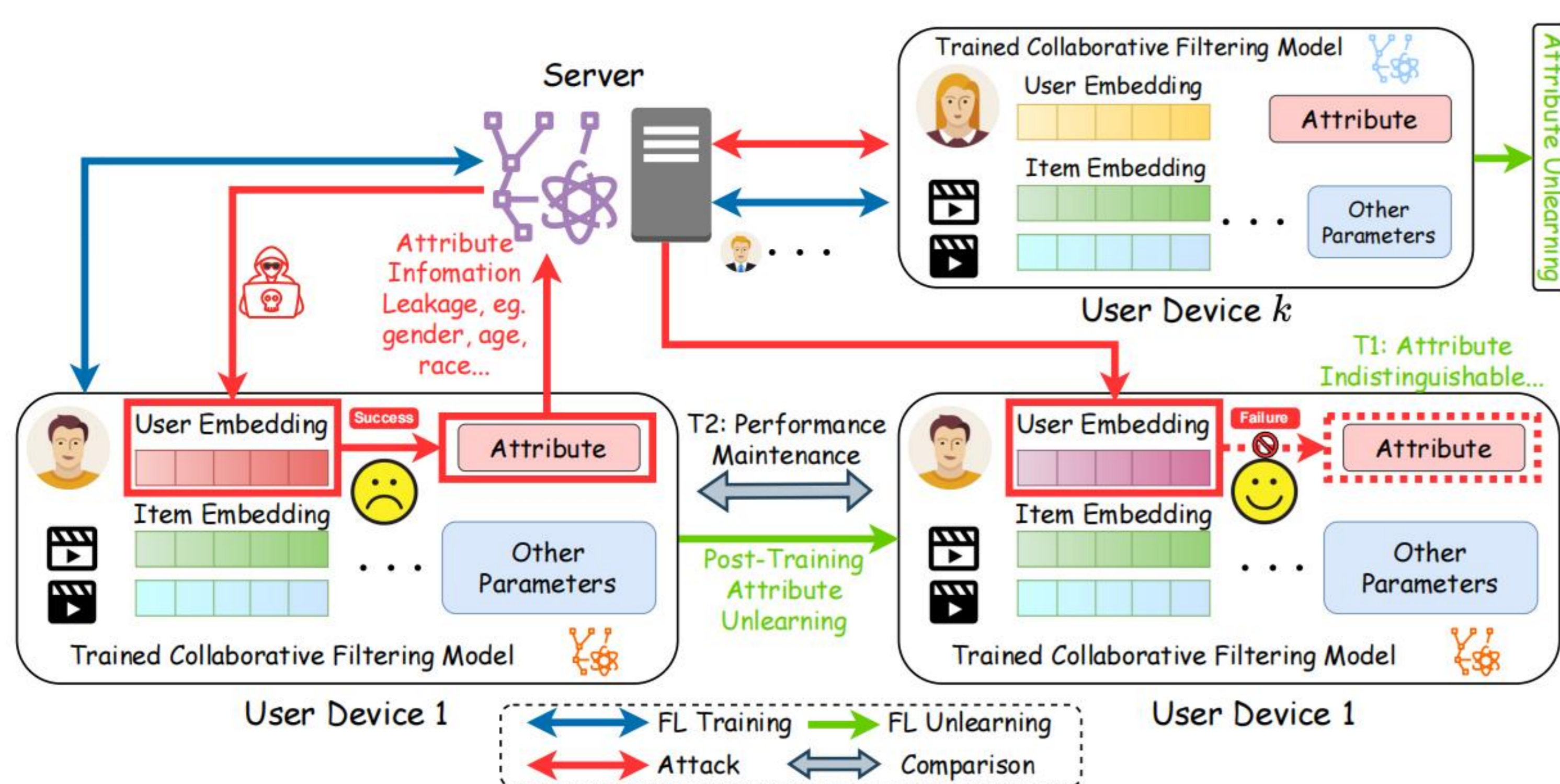
Current privacy mechanisms are usually in-training protection, which rely on network modifications or adversarial training. These in-training attribute-preserving methods are costly, complex, and require prior knowledge of privacy issues, making them less suited to dynamic privacy needs. In real-world scenarios, users' privacy requirements may change over time. Federated clients may want to adjust their privacy settings after training rather than determining them beforehand. This calls for a post-training privacy protection method.

Why attribute unlearning?

- Current machine unlearning methods struggle to decouple sensitive attributes from the model. Attribute unlearning (AU) can effectively remove such attributes, thus defending against AIAs.
- AU, as a post-training method, is more flexible without the modification to the training phase. Also, it is adaptable to dynamic user privacy requests.



System Overview



Objective 1: Private Attribute Unlearning.

Effectively removing the association between user-marked attributes for deletion and user embeddings to prevent privacy leakage.

Minimize the mutual information between the user embedding em'_i and the attributes to be forgotten $au_j \in AU$, defined as:

$$\min_{em'_i} \sum_{au_j \in AU} I(em'_i; au_j)$$

Objective 2: Recommendation Knowledge Retention.

Ensuring the recommendation performance of the FedRecs is maintained post-unlearning.

Maximize the mutual information between em'_i of user u_i and the interaction matrix R_i with item embeddings V_{EM}

$$\max_{em'_i} I(em'_i, V_{EM}; R_i)$$

Compositional Attribute Unlearning

Private Attributes Information Loss.

Aegis approximate the mutual information $\mathcal{L}_{i,j}^{AU}$ of attribute au_j ($|U_{au_j}|$ classes, for class $C_k \in U_{au_j}$, the number of users is $|S_{au_j=C_k}|$) using a variational upper bound based on the KL divergence

$$\mathcal{L}_{i,j}^{AU} = I(em'_i; au_j) \leq \sum_{k=1}^{|U_{au_j}|} \frac{|S_{au_j=C_k}|}{|S_{au_j}|} D_{KL}(q_\phi(em'_i | X_{au_j=C_k}) \parallel p(em'_i))$$

- Computing the user embedding distribution for each class (fitting a Gaussian distribution):

$$\mu_{j,k} = \frac{1}{|S_{au_j=C_k}|} \sum_{em'_i \in S_{au_j=C_k}} em'_i, \quad \Sigma_{j,k} = \frac{1}{|S_{au_j=C_k}|} \sum_{em'_i \in S_{au_j=C_k}} (em'_i - \mu_{j,k})(em'_i - \mu_{j,k})^T$$

$$\mu_{global} = \frac{1}{|S_{au_j}|} \sum_{em'_i \in S_{au_j}} em'_i, \quad \Sigma_{global} = \frac{1}{|S_{au_j}|} \sum_{em'_i \in S_{au_j}} (em'_i - \mu_{global})(em'_i - \mu_{global})^T$$

- Calculating the KL divergence between each class embedding distribution and the global embedding distribution:

$$D_{KL}(N(\mu_{j,k}, \Sigma_{j,k}) \parallel N(\mu_{global}, \Sigma_{global})) = \frac{1}{2} \left[\log \frac{\det(\Sigma_{global})}{\det(\Sigma_{j,k})} - d + \text{Tr}(\Sigma_{global}^{-1} \Sigma_{j,k}) + (\mu_{global} - \mu_{j,k})^T \Sigma_{global}^{-1} (\mu_{global} - \mu_{j,k}) \right]$$

- Calculating the unlearning loss function \mathcal{L}_j^{AU} for au_j as below:

$$\mathcal{L}_j^{AU} = \sum_{k=1}^{|U_{au_j}|} \frac{|S_{au_j=C_k}|}{|S_{au_j}|} D_{KL}(N(\mu_{j,k}, \Sigma_{j,k}) \parallel N(\mu_{global}, \Sigma_{global}))$$

Recommendation Knowledge Retention Loss.

An intuitive approach is to directly use the recommendation loss function \mathcal{L}^{Rec} from the federated training phase. To accelerate the execution process, we only update user embeddings during unlearning:

$$\mathcal{L}^{Rec} = \mathcal{L}_{BPR, BPR, \dots}(s_\psi(f_{\phi,p}(u), f_{\phi,p}(i)), R)$$

We additionally propose the use of a regularization loss \mathcal{L}^{Reg} to restrict the range of user embedding updates:

$$\mathcal{L}^{Reg} = \sum_{i=1}^{|U|} \|em_i - em'_i\|^2 = \sum_{i=1}^{|U|} \sum_{j=1}^d \|em_{i,j} - em'_{i,j}\|^2$$

Attribute Unlearning Fine-tuning Summary.

Pre-Unlearning Stage: We adopt the standard FedRec model to train the recommender system. Then, we have:

- Aegis-Fed (federated protection, data-dependent):

$$\phi, p = \min_{\phi, p} \mathcal{L}^{All} = \min_{\phi, p} (\mathcal{L}^{Rec} + \beta \mathcal{L}^{Reg} + \gamma \sum_{au_j \in AU} \mathcal{L}_{i,j}^{AU})$$

- Aegis-CS (federated protection, data-free):

$$\phi, p = \min_{\phi, p} \mathcal{L}^{All} = \min_{\phi, p} (\beta \mathcal{L}^{Reg} + \gamma \sum_{au_j \in AU} \mathcal{L}_{i,j}^{AU})$$

Evaluation Results

Evaluation Metrics:

1. AIA accuracy.
2. Hit Ratio (HR@5,10).
3. Normalized Discounted Cumulative Gain (NDCG @5,10).
4. Running Time.

Attack Setting

1. Threat Model: MLP/XGBoost.
2. Gray-Box Strategy: the attacker cannot access all model parameters but can access some user embedding vectors.

Attribute-wise Privacy Performance against AIA.

Dataset		MovieLens-100K			MovieLens-1M			ModCloth	Last.FM-1K			
Sensitive Attributes		Gender	Age	Occupation	Gender	Age	Occupation	Body Shape	Gender	Age	Location	
XGBoost Attacker	DD	Original	0.7143	0.2804	0.1490	0.8487	0.3526	0.1192	0.7419	0.5989	0.4828	0.5604
		Aegis-Fed	0.5703	0.2222	0.1058	0.5968	0.1798	0.0589	0.5316	0.4798	0.2778	0.3586
		UC-FedRec	0.6772	0.2857	0.2011	0.7280	0.3220	0.1543	0.7325	0.5606	0.2980	0.4899
	DF	Aegis-CS	0.5450	0.2116	0.0794	0.6450	0.2064	0.0766	0.5709	0.4506	0.2929	0.3766
		U2U-R	0.9921	0.5947	0.6980	0.9997	0.9998	0.9940	0.9999	0.9999	0.8690	0.9518
		D2D-R	0.5834	0.2464	0.1020	0.6821	0.2941	0.0923	0.7799	0.4746	0.4585	0.4876
MLP Attacker	DD	Original	0.7105	0.3386	0.0954	0.7310	0.3470	0.1159	0.7654	0.6102	0.5350	0.5604
		Aegis-Fed	0.6541	0.1376	0.0794	0.5625	0.1598	0.0762	0.6602	0.5556	0.1162	0.2828
		UC-FedRec	0.6085	0.1640	0.0582	0.3654	0.1836	0.0263	0.6784	0.4242	0.1364	0.2171
	DF	Aegis-CS	0.6224	0.1693	0.0688	0.6032	0.1821	0.0389	0.6462	0.5253	0.0690	0.2677
		U2U-R	0.6931	0.2646	0.0370	0.7193	0.0704	0.0985	0.7107	0.5650	0.3536	0.4088
		D2D-R	0.6720	0.2434	0.0529	0.6987	0.2359	0.0298	0.6728	0.5480	0.3103	0.2802
Random Attacker		0.5000	0.1429	0.0476	0.5000	0.1429	0.0476	0.5000	0.5000	0.1469	0.1667	

Recommendation Performance.

Datasets	Methods	Utility Metrics				
		NDCG@5	NDCG@10	HR@5	HR@10	
MovieLens-100K	Original	0.7452	0.7080	0.7234	0.6804	
	DD	Aegis-Fed	0.7632	0.6905	0.7321	0.6959
		UC-FedRec	0.6959	0.6452	0.7032	0.6698
	DF	Aegis-CS	0.7209	0.6891	0.6944	0.6571
		U2U-R	0.7151	0.6854	0.7040	0.6663
		D2D-R	0.7194	0.6862	0.7129	0.6693
MovieLens-1M	Original	0.6992	0.6901	0.6958	0.6843	
	DD	Aegis-Fed	0.6332	0.6262	0.6292	0.6211
		UC-FedRec	0.6320	0.6525	0.6280	0.6204
	DF	Aegis-CS	0.6631	0.6707	0.6641	0.6604
		U2U-R	0.6481	0.6156	0.6420	0.5965
		D2D-R	0.6928	0.6623	0.6388	0.6929
ModCloth	Original	0.6077	0.6079	0.6047	0.6115	
	DD	Aegis-Fed	0.6044	0.6071	0.6352	0.6386
		UC-FedRec	0.5664	0.5551	0.5622	0.5542
	DF	Aegis-CS	0.5969	0.5814	0.5563	0.5530
		U2U-R	0.5605	0.5437	0.5482	0.5319
		D2D-R	0.5854	0.5694	0.5653	0.5561
Last.FM-1K	Original	0.5724	0.5665	0.5806	0.5680	
	DD	Aegis-Fed	0.5888	0.5939	0.5962	0.5999
		UC-FedRec	0.5182	0.5190	0.5243	0.5229
	DF	Aegis-CS	0.5446	0.5726	0.5442	0.5052
		U2U-R	0.5182	0.5190	0.5244	0.5229
		D2D-R	0.5282	0.5975	0.5362	0.5490

Efficiency.

Time(s)	Aegis-Fed	UC-FedRec	Aegis-CS	U2U-R	D2D-R
ML-100K	392.04	941.74	10.91	9.80	5.43
ML-1M	4742.96	8223.53	37.66	109.68	54.45
ModCloth	571.22	981.43	165.18	167.37	87.35
Last.FM-1K	65085.79	163129.51	8.88	12.15	5.68

Comprehensive evaluations demonstrate that Aegis effectively safeguards user privacy (with AIA accuracy closer to that of a random attacker) while maintaining high-quality recommendations (without great performance drop), in an efficient post-training manner (with low running time).

