Shadow-Free Membership Inference Attacks: Recommender Systems Are More Vulnerable Than You Thought

Xiaoxiao Chi[1], **Xuyun Zhang**[1*], **Yan Wang**[1], **Lianyong Qi**[2], **Amin Beheshti**[1], **Xiaolong Xu**[3], **Kim-Kwang Raymond Choo**[4], **Shuo Wang**[5], **Hongsheng Hu**[6*]
[1]Macquarie University
[2]China University of Petroleum (East China)
[3]Nanjing University of Information Science and Technology
[4]The University of Texas at San Antonio
[5]Shanghai Jiao Tong University
[6]CSIRO's Data61

Abstract

Recommender systems have been successfully applied in many applications. Nonetheless, recent studies demonstrate that recommender systems are vulnerable to membership inference attacks (MIAs), leading to the leakage of users' membership privacy. However, existing MIAs relying on shadow training suffer a large performance drop when the attacker lacks knowledge of the training data distribution and the model architecture of the target recommender system. To better understand the privacy risks of recommender systems, we propose shadow-free MIAs that directly leverage a user's recommendations for membership inference. Without shadow training, the proposed attack can conduct MIAs efficiently and effectively under a practice scenario where the attacker is given only black-box access to the target recommender system. The proposed attack leverages an intuition that the recommender system personalizes a user's recommendations if his historical interactions are used by it. Thus, an attacker can infer membership privacy by determining whether the recommendations are more similar to the interactions or the general popular items. We conduct extensive experiments on benchmark datasets across various recommender systems. Remarkably, our attack achieves far better attack accuracy with low false positive rates than baselines while with a much lower computational cost.

1 Introduction

Recommender systems aim to accurately predict and suggest items or contents for users, which are widely applied in many real-world applications [Zhang *et al.*, 2019], such as e-commerce sites [Zhou *et al.*, 2018], healthcare domains [Narducci *et al.*, 2015], and social platforms [Tang *et al.*, 2016; Wu *et al.*, 2019; Fan *et al.*, 2019]. The success of recommender systems is largely attributed to the increasing availability of large-scale data generated by or associated with end users. The data often contain user profiles or behavioral information like age, gender, and shopping preference, thereby requiring strong protection on user privacy in terms of many recently issued laws and regulations like GDPR [Rosen, 2011] and CCPA [Pardau, 2018]. However, recent studies [Zhang *et al.*, 2021; Wang *et al.*, 2022; Zhu *et al.*, 2023] demonstrate that recommender systems are vulnerable to membership inference attacks (MIAs) [Shokri *et al.*, 2017], where an attacker can infer the membership privacy of a user, i.e., distinguish member users whose data was used for training the recommender system from non-member users of the model. MIAs can directly reveal the privacy of a user in recommender systems. For example, if an attacker identifies that a user's data has been used for training a healthcare recommender system for treatment plans, the attacker can infer the user is a patient with a high chance. In addition, MIAs can be the foundations of other types of attacks, e.g., data extraction attacks [Carlini *et al.*, 2019; Carlini *et al.*, 2021]. Because of such abilities, MIAs have been widely used for measuring the privacy risks of machine learning models [Carlini *et al.*, 2022; Hu *et al.*, 2022].

To implement MIAs, a common approach is *shadow training*, where the attacker trains a shadow model to mimic the behavior of the target model. Because the shadow model is trained by the attacker, the attacker can collect the features of members and non-members of the shadow model, which can be used for training a binary classifier as the attack model. Because the shadow model mimics the target model, the attack model trained on the shadow model will also work on the target model, which is often referred to as the attack transferability [Salem *et al.*, 2019]. Following the pipeline of shadow training, existing studies on MIAs targeting recommender systems [Zhang *et al.*, 2021; Wang *et al.*, 2022; Zhu *et al.*, 2023] are highly effective in inferring whether an individual's data was used to train a recommender system or not, e.g., the work in [Zhang *et al.*, 2021] shows that MIAs with the use of shadow training can achieve attack accuracy near 100% against an item-based collaborative filtering rec-

*Corresponding Author. Contact xuyun.zhang@mq.edu.au or Hongsheng.Hu@data61.csiro.au

ommender system trained on the Movielens-1M dataset. As a pre-requisite of shadow training, existing MIAs on recommender systems have a key assumption that the attacker owns the prior knowledge about the training data distribution and the model architecture of the target model to ensure the attack transferability. [Zhang *et al.*, 2021; Wang *et al.*, 2022; Zhu *et al.*, 2023]. With this assumption, the shadow model can be expected to behave similarly to the target model. However, it is often difficult for an attacker to obtain the prior knowledge in practice and the assumption fails to hold, since recommender systems are usually deployed under MLaaS (Machine Learning as a Service) environments [Ribeiro *et al.*, 2015] and only black-box access is available to the public. Besides, a shadow dataset from the same distribution of the training dataset is very difficult to satisfy completely. While existing works [Zhang *et al.*, 2021; Wang *et al.*, 2022; Zhu *et al.*, 2023] claim their abilities to generate a shadow dataset through querying the target model or using marginal distributions of training data, they often simply set a part of training dataset aside as the shadow dataset in empirical evaluations. In addition to this limitation, MIAs with shadow training are computationally expensive, requiring considerable computational resources for training both the shadow model and the attack model. Therefore, how to address these two drawbacks is still a challenge in existing MIAs with shadow training.

From a defence perspective, the above-mentioned impractical assumption may give a sense of security in recommender systems: They can stay safe from MIAs as long as the attacker does not have the assumed prior attack knowledge. Indeed, existing works [Zhang *et al.*, 2021; Wang *et al.*, 2022; Zhu *et al.*, 2023] have tried to train a shadow model without the prior attack knowledge. Instead, they make use of a different shadow model architecture and a shadow dataset from a different distribution, but the resultant attack model suffers from a large performance drop, sometimes even to a level of randomly guessing. This demonstrates failure to have the assumption hold will disable MIAs with shadow training. However, this sense of security is false as existing MIAs on recommender systems fail to take the special aspects of recommender systems into consideration. Recommender systems usually recommend personalized items to members, as their historical interactions were used for training the model. Such personalized items, through meticulously chosen by the recommender system, are similar to members' historical interactions. However, for non-members, recommender systems usually recommend generally popular items, because the model has not seen their historical interactions [Sedhain *et al.*, 2014].

In regard of this, in this paper we propose shadow-free MIAs without any process of shadow training with only black-box access to the model. The attack intuition is to examine whether a user's recommendations are more similar to his historical interactions or general popular items. If the recommendations are more similar to historical interactions, the attacker can infer the user as a member, and infer the user as a non-member otherwise. The challenge here is how to obtain general popular items of a recommender system, which serves as an important reference for the similarity comparison. To solve this challenge, we skillfully leverage the characteristics of recommender system scenarios. Specifically, an attacker can generate an empty user account without historical interactions with the target recommender system. Then, the attacker can collect the recommendations of the empty user and consider them as general popular items. This implementation is easy for the attacker to achieve in practice, e.g., creating a new account in Amazon. Being lightweight, the newly proposed attacks can efficiently and effectively conduct MIAs against recommender systems.

Our contribution is summarized as follows:

- This paper is the *first* to investigate shadow-free MIAs against recommender systems. The proposed new attack is lightweight and can effectively infer the membership privacy of a user with only black-box access to the recommender system.
- Extensive experiments are conducted on three benchmark datasets across various recommender systems and compared with representative baseline attacks. Experimental results demonstrate that the newly proposed attacks achieve far better performance than baselines in terms of attack effectiveness, reliability, and attack efficiency.
- The source code of the shadow-free MIAs is released at https://github.com/XiaoxiaoChi-code/shadow-free-MIAs.git, which creates a new tool for measuring the privacy vulnerability of recommender systems and sheds light on the design of future defense methods.

2 Related Work

Recommender Systems. Recommender systems enrich user experiences by predicting and suggesting items within a vast array of content. Among various recommendation algorithms [Burke, 2002; Chen *et al.*, 2017], traditional collaborative filtering recommendation algorithm [Koren *et al.*, 2009] is the main stream, which aims to recommend items to users based on their preferences and behaviors of other users with similar tastes. In recent years, deep learning techniques have been widely applied to recommender systems. Advanced deep learning based recommender systems leverage neural networks to model complex patterns and representations of user-item interactions. Techniques such as autoencoders [Sedhain *et al.*, 2015], neural collaborative filtering [He *et al.*, 2017], recurrent neural networks [Hidasi *et al.*, 2015], and long short-term memory networks [Liu *et al.*, 2018; Zhou *et al.*, 2019] have demonstrated significant success in improving recommendation accuracy and addressing challenges associated with sparse and high-dimensional data.

Membership Inference Attacks. MIAs aim to infer whether a data sample was used to train a target model or not. The work [Shokri *et al.*, 2017] firstly investigates MIAs on classification models. Later works [Hayes *et al.*, 2017; Song and Shmatikov, 2019; He *et al.*, 2020; He *et al.*, 2021] further investigate the feasibility of MIAs on other types of models such as image generative and segmentation models. Given the simplicity of the definition, MIAs have been considered as a standard metric for measuring the privacy of machine learning models [Carlini *et al.*, 2022; Ye *et al.*, 2022;

Song and Mittal, 2021]. A few works [Zhang *et al.*, 2021; Zhu *et al.*, 2023; Wang *et al.*, 2022] have investigated the membership privacy risks on recommender systems and demonstrated that MIAs are less effective if an attacker does not know the prior knowledge of the training data distribution and the model architecture of target recommender system. This phenomenon is reasonable because existing MIAs rely on shadow training. Once the shadow recommender system is not similar enough to the target one, the attack model built on the shadow recommender system cannot transfer well to the target one. In addition, shadow training is usually associated with high computational costs for training the shadow and attack models. In this paper, we fulfill this research gap by proposing shadow-free MIAs, which can efficiently and effectively conduct the MIAs without shadow training.

3 Methodology

In this section, we first introduce the threat model of MIAs in recommender systems. Then, we introduce shadow-based MIAs and analyze their limitations, which serve as a motivation for proposing more powerful and practical MIAs. Last, we introduce our proposed method for shadow-free MIAs.

3.1 Threat Model

In this paper, we study MIAs under the black-box settings, i.e., we assume an attacker can only query the target recommender system and obtain the recommended items. Following previous works [Zhang *et al.*, 2021; Wang *et al.*, 2022], we assume the attacker has a dataset that contains users' ratings of items. This dataset can be obtained via generative methods or crawled from the internet [Zhu *et al.*, 2023], and it is used for generating item features using matrix factorization [Koren *et al.*, 2009]. Unlike the existing works in [Zhang *et al.*, 2021; Wang *et al.*, 2022], we do not assume the attacker has a shadow dataset that comes from the same distribution as the training dataset of the recommender system. In addition, we do not assume the attacker has knowledge of the model architecture of the recommender system. In contrast, the availability of such knowledge is a key factor for the success of MIAs in existing works [Zhang *et al.*, 2021; Wang *et al.*, 2022] following shadow training.

3.2 Shadow-based MIAs and Their Limitations

Notations. Let $\boldsymbol{M}^{p\times q}$ be a user-item matrix that contains p users' ratings of q items. Using the matrix factorization [Koren *et al.*, 2009] technique, we can divide $\boldsymbol{M}^{p\times q}$ into the product of two lower dimensional matrices:

$$\hat{\boldsymbol{M}}^{p\times q} = \boldsymbol{H}^{p\times l} \cdot \boldsymbol{W}^{l\times q}, \tag{1}$$

by minimizing the following loss function:

$$\min ||\boldsymbol{M}^{p\times q} - \hat{\boldsymbol{M}}^{p\times q}||_2, \tag{2}$$

where $||\cdot||_2$ is the l-2 norm. $\boldsymbol{H}$ contains the user's latent factors, $\boldsymbol{W}^{\mathrm{T}}$ contains item's latent factors. We denote $\boldsymbol{W}^{\mathrm{T}} = (\boldsymbol{w}_1; \cdots; \boldsymbol{w}_q)$, where each $\boldsymbol{w}_i$ is a l-dimensional vector and represents the feature of an item. Let $\boldsymbol{x} = [x_1, \cdots, x_m]$ be m historical interactions of a user, where each x_i is an item. A recommender system is a function $f(\cdot)$ that takes as input the historial interactions of a user and outputs n recommended items $\boldsymbol{Y} = f(\boldsymbol{x})$ to the user. Specifically, $\boldsymbol{Y} = [y_1, \cdots, y_n]$ is a n-dimensional vector and each y_i represents a recommended item.

Existing Shadow-based Membership Inference. Shadow-based MIAs aim to train a binary classifier $h(\cdot)$ that takes as an input a user's feature vector $\boldsymbol{v}$ and outputs 0 or 1:

$$h : \boldsymbol{V} \rightarrow \{0, 1\}, \tag{3}$$

where $\boldsymbol{V}$ represents users' feature vector space and $\boldsymbol{v} \in \boldsymbol{V}$, 0 represents that the attack classifier predicts the user as a non-member, and 1 as a member. To train the binary attack classifier, the attacker requires to obtain the features of member users and non-member users.

Existing shadow-based MIAs assume the attacker can have a shadow dataset $\mathcal{D}_s$ that comes from the same distribution of the training data of the target recommender system. In addition, the attacker is assumed to know the model architecture $\mathcal{A}$ of the target recommender system. There are three steps in the shadow-based MIAs: *i)* The attacker splits $\mathcal{D}_s$ into two disjoint datasets: $\mathcal{D}_s^{\text{train}}$ and $\mathcal{D}_s^{\text{test}}$. Following the same model architecture of $\mathcal{A}$, the attacker trains a shadow model $f_s(\cdot)$ on $\mathcal{D}_s^{\text{train}}$ to mimic the behavior of the target model; *ii)* After the training of $f_s(\cdot)$, for each user's historical interactions $\boldsymbol{x} \in \mathcal{D}_s^{\text{train}}$, the attacker queries the shadow model and records the corresponding recommended items $\boldsymbol{Y}_{\text{train}} = [y_1, \cdots, y_n]$. Since the attacker has $\boldsymbol{W}^{\mathrm{T}}$, the attacker can obtain a feature vector as:

$$\boldsymbol{v} = \frac{1}{m}\sum_{m=1}^{m} \boldsymbol{w}_{x_m} - \frac{1}{n}\sum_{n=1}^{n} \boldsymbol{w}_{y_n}, \tag{4}$$

where $\boldsymbol{v}$ is a l-dimensional vector and $\boldsymbol{w}_i$ is the corresponding feature vector of the item i. The attacker considers such a vector as the feature vector of a user since it encodes information on both recommendations and historical interactions. Based on $\boldsymbol{x} \in \mathcal{D}_s^{\text{train}}$ and $\boldsymbol{x} \in \mathcal{D}_s^{\text{test}}$, the attacker can obtain feature vectors of member and non-member users by querying the shadow model; *iii)* Since the attacker has collected features of users in the previous step, the attacker now can train a binary classifier using standard machine learning training procedures. After training, the binary classifier works as an attack model for predicting the membership status of a target user of the target recommender system.

Limitations of Existing Attacks. There are two limitations of existing shadow-based MIAs. First, shadow-based MIAs can not work effectively when the target dataset is not available or when the model architecture of the target recommender system is unknown. This is because, under this scenario, the shadow model cannot mimic the behavior of the target model well. Second, constructing the attack model is relatively computationally expensive because the attacker is required to train a shadow model and an attack model. When both of the models are complex, following the same architecture, the attacker requires large computational resources for training the shadow model. In Section 4.2, we demonstrate that in some cases where the target model is a deep learning-based recommender system, it takes a long time to train the attack model. To solve the two limitations, we propose shadow-free MIAs: an attacker is able to efficiently and

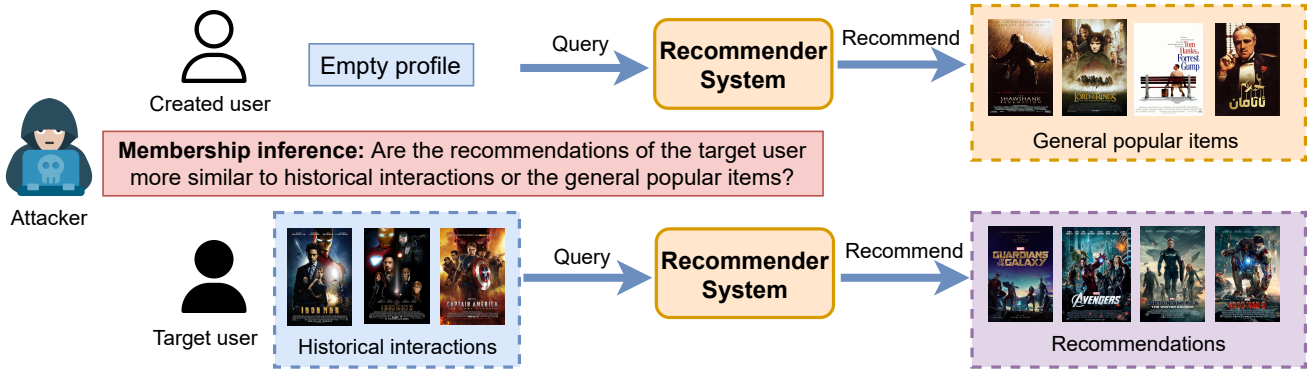

Figure 1: An overview of shadow-free MIAs. The attacker creates a user with an empty profile to obtain the general popular items of the recommender system. For a target user, the attacker examines whether the recommendations of the target user are more similar to his historical interactions or the general popular items to determine the membership status of the target user.

effectively conduct membership inference without the training of a shadow model and the requirements of the target dataset or knowledge of the target recommender system.

3.3 Shadow-free Membership Inference

Key Intuition. Figure 1 shows an overview of shadow-free MIAs. The key of the attack is to compare the recommendations of the target user to his historical interactions and the general popular items. A key observation is leveraged to determine the membership status of a user: A member user's recommendations provided by the recommender system are similar to his interactions because such interactions are used by the recommender system to find relevant items. As depicted in Figure 1, if a user likes Marvel movies and his data was used by the recommender system, it is highly likely to recommend other Marvel movies to the user. This makes the recommendations of a member user much more similar to the interactions than the general popular items recommended by the recommender system, e.g., general high-rated movies depicted in Figure 1. Thus, the attacker can infer a user as a member if the recommendations are more similar to the interactions than the general popular items, and infer the user as a non-member otherwise.

Formally, there are three steps for conducting shadow-free MIAs, described as follows:

i) Creating an Empty User. To obtain the general popular items of the recommender system, the attacker first creates a user account with no interactions with the recommender system. This is not difficult to achieve in practice, e.g., the attacker can easily register a new account in IMDB. Because the attacker has black-box access to the recommender system, the attacker can obtain n popular items using the newly created user account. We define these popular items as:

$$\boldsymbol{Y}_p = [y_1, \cdots, y_n]. \tag{5}$$

Using the item's latent matrix $\boldsymbol{W}^{\mathrm{T}}$ and each y_i in $\boldsymbol{Y}_p$, the attacker can obtain a feature vector of $\boldsymbol{Y}_p$:

$$\boldsymbol{v}_p = \frac{1}{n}\sum_{n=1}^{n} \boldsymbol{w}_{y_n}. \tag{6}$$

ii) Query the Recommender System. For a target user with historical interactions $\boldsymbol{X} = [x_1, \cdots, x_m]$, the attacker queries the target model and obtains n recommended items. We define these recommendations as:

$$\boldsymbol{Y}_t = [y_1, \cdots, y_n]. \tag{7}$$

Using the item's latent matrix $\boldsymbol{W}^{\mathrm{T}}$, each x_i in $\boldsymbol{X}$, and each y_i in $\boldsymbol{Y}_t$, the attacker can obtain a feature vector of $\boldsymbol{X}$ and $\boldsymbol{Y}_t$, respectively:

$$\boldsymbol{v}_x = \frac{1}{m}\sum_{m=1}^{m} \boldsymbol{w}_{x_m}, \tag{8}$$

$$\boldsymbol{v}_t = \frac{1}{n}\sum_{n=1}^{n} \boldsymbol{w}_{y_n}. \tag{9}$$

iii) Infer the Membership Privacy. To determine the membership status of the target user, the attacker needs to determine whether the recommendations are more similar to the interactions or the general popular items. This can be done by calculating the corresponding similarities and comparing them. Given the feature vectors $\boldsymbol{v}_p$, $\boldsymbol{v}_x$, and $\boldsymbol{v}_t$, the attacker calculates two distances:

$$\alpha_1 = ||\boldsymbol{v}_p - \boldsymbol{v}_t||_2, \tag{10}$$

$$\alpha_2 = ||\boldsymbol{v}_x - \boldsymbol{v}_t||_2. \tag{11}$$

In the context of recommender systems, α_1 represents how close the recommendations are to the general popular items, while α_2 represents how similar the recommendations are to the interactions of the target user. In general, a larger α represents a smaller similarity. If $\alpha_1 > \alpha_2$, indicating that the recommendations are more similar to the interactions of the target user, the attacker considers the user as a member. Otherwise, the attacker considers the user as a non-member. Formally, the attack $\mathcal{M}(\cdot,\cdot)$ is defined as follows:

$$\mathcal{M}(\alpha_1, \alpha_2) = \begin{cases} 1 & \text{if } \alpha_1 > \alpha_2, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Essentially, our shadow-free MIA is metric-based MIA, analogous to metric-based MIAs in the context of classification models [Yeom *et al.*, 2018; Song and Mittal, 2021]

Attacks	Target recommender systems														
	ICF			NCF			BERT4Rec			Caser			GRU4Rec		
	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR
SF-MIAs (Ours)	**0.793**	0.611	**0.025**	**0.968**	0.960	**0.025**	0.808	0.675	**0.059**	**0.986**	0.997	**0.025**	**0.983**	0.999	0.033
ST-MIA	BT 0.500	0.999	1.000	BT 0.499	0.997	1.000	GA 0.403	0.451	0.645	BA 0.500	**1.000**	1.000	CA 0.503	**1.000**	0.995
	NT 0.500	**1.000**	1.000	IT 0.502	**1.000**	0.996	CA 0.493	0.982	0.997	GA 0.670	0.999	0.658	BA 0.500	**1.000**	1.000
DL-MIA	BT 0.582	**1.000**	0.839	BT 0.502	**1.000**	1.000	GA 0.340	0.001	0.199	BA 0.930	0.860	0.124	CA 0.950	**1.000**	**0.032**
	NT 0.504	**1.000**	0.991	IT 0.502	**1.000**	1.000	CA **0.868**	**1.000**	0.266	GA 0.984	**1.000**	0.032	BA 0.884	**1.000**	0.233

Table 1: Attack accuracy, TPR, and FPR of the shadow-free MIAs (SF-MIAs) and the two attack baselines across five recommender systems on the MovieLens-1M dataset.

where prediction vectors of data records are used for calculating metrics and the calculated metrics are then compared with a preset threshold to determine the membership status of data records. However, different from these metric-based MIAs, the metric (i.e., comparing α_1 with α_2) in our attacks is self-adaptive as it is calculated in a user-specific manner. Thus, in our attacks, preset threshold is unnecessary. The self-adaptiveness is a big advantage of our method compared to other metric-based approaches where determining the threshold value is a non-trivial job for an attacker. The technical innovations of shadow-free MIAs bring substantial benefits including a simplified methodology design, wide applicability with a practical black-box assumption, improved attack accuracy, and low computation cost.

4 Experiments

4.1 Experimental Setup

We conduct extensive experiments on three benchmark datasets across five different recommender systems, which include traditional recommender system and advanced deep learning based ones. Due to page limits, the detailed description of datasets, dataset partition, main parameter settings in recommender systems, detailed introduction of baselines, baseline attack setting descriptions, and facilities utilized for experiments are available in Appendix A.1[1].

Datasets. In the experiments, three benchmark datasets are leveraged: MovieLens-1M [Harper and Konstan, 2015], Amazon Beauty [McAuley *et al.*, 2015], and Ta-feng [2]. All these datasets are benchmark datasets for evaluating the performance of recommender systems.

Recommender Systems. We select five representative recommender systems to comprehensively evaluate our proposed attacks. These recommender systems including the traditional recommender system of the Item-based Collaborative Filtering (ICF) [Sarwar *et al.*, 2001], as well as the advanced deep learning based ones of the Neural Collaborative Filtering (NCF) [He *et al.*, 2017], BERT4Rec [Sun *et al.*, 2019], Caser [Tang and Wang, 2018], and GRU4Rec [Hidasi *et al.*, 2015]. Following previous works [Zhang *et al.*, 2021; Wang *et al.*, 2022], personalized recommendation lists are generated for existing users via recommendation algorithms of the recommender systems. For new users, due to a lack of their data, recommender systems recommend the most popular items to them.

Baselines. We compare the proposed shadow-free MIAs (SF-MIAs) with two state-of-the-art attack methods: shadow-based MIAs (ST-MIAs) [Zhang *et al.*, 2021] and Debiasing learning for MIAs (DL-MIAs) [Wang *et al.*, 2022]. The two baselines in our experiments are with the same black-box assumption as our approach.

Evaluation Metrics. We evaluate MIAs from two perspectives: effectiveness and efficiency. In terms of effectiveness, since membership inference is a binary classification problem, we use accuracy to evaluate the attack performance, which is one of the most widely used metrics in existing studies of MIAs [Hu *et al.*, 2022]. In addition, as suggested by [Carlini *et al.*, 2022] that a powerful and reliable MIA should have a high true positive rate (TPR) at a low false positive rate (FPR), we report TPR and FPR of our attacks and the two baselines. An attack is considered to be reliable if it can achieve a high true positive rate at a very low false positive rate. In terms of efficiency, we record the overall computational time of the attack model of different approaches. An attack that has a short overall computational time is considered as efficient.

Attack Settings. For baseline methods, we run experiments against combinations of different shadow models and shadow datasets to achieve a comprehensive and fair evaluation.

4.2 Efficacy of The Proposed Attack

Attack Effectiveness. Table 1, Table 2, and Table 3 show attack accuracy, TPR, and FPR of our attack and the two baseline attacks. Based on the experimental results, we can observe that: (*i*) In general, our proposed shadow-free MIAs have very high attack accuracy. The accuracy scores of our method across all attack settings are above 60%. In addition, the accuracy of our proposed attack is above 80% across 70% of attack settings, demonstrating the vulnerabilities of recommender systems in most settings. In some cases, our attacks achieve perfect performance with an accuracy close to 100%. For example, in Table 1, the highest accuracy of our method can be achieved at 98.6% when the Caser recommender system is trained on MovieLens-1M; (*ii*) Our shadow-free MIAs consistently outperform the baselines in all experimental settings in Table 1 with respect to accuracy. For Table 2 and Table 3, the accuracy of our attacks is higher than the two

[1]Please refer to the version of this paper with Appendix in arXiv.

[2]https://www.kaggle.com/datasets/chiranjivdas09/ta-feng-grocery-dataset

Attacks	Target recommender systems														
	ICF			NCF			BERT4Rec			Caser			GRU4Rec		
	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR
SF-MIAs (Ours)	**0.815**	0.630	**0.000**	**0.620**	0.241	**0.000**	**0.693**	0.386	**0.000**	0.727	0.453	**0.000**	**0.923**	0.845	**0.000**
ST-MIA	BM 0.520	0.517	0.478	IM 0.571	0.602	0.461	CM 0.543	0.117	0.031	BM 0.584	0.472	0.304	CM 0.890	0.796	0.016
	BT 0.330	**0.636**	0.976	BT 0.414	0.804	0.976	GM 0.553	0.153	0.048	GM 0.736	0.510	0.038	BM 0.673	0.633	0.287
DL-MIA	BM 0.468	0.280	0.343	IM 0.503	**0.938**	0.933	CM 0.509	0.922	0.904	BM 0.519	**0.991**	0.954	CM 0.503	**0.913**	0.894
	BT 0.498	0.007	0.010	BT 0.505	0.014	0.005	GM 0.572	**0.958**	0.816	GM **0.916**	0.951	0.120	BM 0.346	0.314	0.623

Table 2: Attack accuracy, TPR, and FPR of the shadow-free MIAs (SF-MIAs) and the two attack baselines across five recommender systems on the Amazon Beauty dataset.

Attacks	Target recommender systems														
	ICF			NCF			BERT4Rec			Caser			GRU4Rec		
	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR	Accuracy	TPR	FPR
SF-MIAs (Ours)	**0.981**	0.961	**0.000**	0.856	0.713	**0.000**	**0.645**	0.210	**0.000**	**0.977**	0.954	**0.000**	**0.991**	0.983	**0.000**
ST-MIA	CA 0.950	0.998	0.098	GA 0.875	0.916	0.165	CA 0.579	0.244	0.087	GA 0.897	**0.959**	0.165	BA 0.572	**1.000**	0.855
	BM 0.897	0.829	0.036	BM 0.673	0.381	0.036	GM 0.495	0.002	0.013	NM 0.522	0.947	0.902	BM 0.922	0.875	0.031
DL-MIA	CA 0.500	**1.000**	1.000	GA 0.939	0.991	0.113	CA 0.499	**0.999**	0.999	GA 0.889	0.895	0.117	BA 0.939	0.904	0.105
	BM 0.682	0.724	0.360	BM **0.947**	**0.993**	0.099	GM 0.449	0.005	0.006	NM 0.703	0.682	0.277	BM 0.696	0.679	0.288

Table 3: Attack accuracy, TPR, and FPR of the shadow-free MIAs (SF-MIAs) and the two attack baselines across five recommender systems on the Ta-feng dataset.

Method	Shadow-free MIAs	ST-MIA	DL-MIA
Time cost (avg)	3.7s	128.8s ($\approx 35\times$)	9,760s ($\approx 2,637\times$)

Table 4: The computational cost of shadow-free MIAs and baseline attacks, measured by seconds. ($T\times$) represents T times faster of the shadow-free MIAs than the baselines.

baselines in almost 80% experimental settings. In addition, in some cases, our attack can achieve near-perfect performance, while the baselines have performance close to random guess. For example, in Table 1, when NCF is trained on MovieLens-1M, the accuracy of our attacks is 96.8%, while the attack accuracy of baselines is nearly close to 50%. (*iii*) MIAs based on shadow training (i.e., the two baselines) have severe limitations when the attacker does not know the training data distribution and the model architecture of the target recommender system. For example, in Table 1, the accuracy of ST-MIA is close to 50% (random guess) in almost 80% experimental settings. For DL-MIAs, although attack accuracy is improved in most cases, they still perform worse than our attacks.

Takeaway 1. *The proposed shadow-free MIAs are more effective than the two baselines. In most cases, the proposed attacks achieve an attack accuracy higher than 0.8, while the baseline attacks can only achieve performance close to random guess with an accuracy of around 0.5.*

Attack Reliability. As suggested by [Carlini *et al.*, 2022], a powerful and reliable MIA should have a high TPR at a low FPR. From the experimental results in Table 1, Table 2, and Table 3, our attack can achieve high TPRs at low FPRs close to 0. For instance, in Table 1, when GRU4Rec is trained on MovieLens-1M, our attack achieves a TPR of 99.9% with an FPR of 3.3%. In Table 2, when GRU4Rec is trained on Amazon Beauty, our attack achieves a TPR of 84.5% with an FPR of 0, demonstrating that no non-member users are mistakenly predicted as member users. In Table 3, when the target recommender system is ICF, NCF, Caser, and GRU4Rec, the TPR of our attack is 96.1%, 71.3%, 95.4%, and 98.3% with FPRs of 0. In contrast, the FPR values tend to be high for baselines. For instance, in Table 2, when target recommender NCF is trained on the target dataset Amazon Beauty and shadow recommender BERT4Rec is trained on Ta-feng, ST-MIA attack achieves a TPR of 80.4% with an FPR of 97.6%. When the target recommender model of NCF is trained on the target dataset Amazon Beauty, the shadow model using ICF trained on MovieLens-1M, DL-MIA attack achieves a TPR of 93.8% with an FPR of 93.3%. The experimental results demonstrate that our proposed SF-MIA is more powerful and reliable.

Takeaway 2. *The shadow-free MIAs are more reliable than baseline attacks. In most cases, the proposed attacks achieve high TPRs with low FPRs close to 0, while the baselines have high FPRs in most cases.*

Attack Efficiency. To compare the efficiency of attack methods, we record the overall computational time cost of different attack methods. For each method, we select five experimental settings to run, and the average time cost is calculated. For instance, we select "(BERT., Ama.)", "(BERT., Mov.)", "(BERT., Ta.)", "(Cas., Ama.)", and "(Cas., Mov.)" settings to calculate the time cost for SF-MIA. For ST-MIA and DL-MIA, we select settings "IM" from Table 2, settings "NT", "IT" from Table 1, and settings "NM" and "BA" from Table 3 as experimental settings to calculate time cost. As we can see in Table 4, our attack method takes only about 3.7 seconds to complete the attack. In contrast, ST-MIA takes almost 128.8 seconds to implement the attack, and DL-MIA is rather expensive in terms of time cost, with around 9,760 seconds. This validates that our attacks are much more efficient

compared with baseline attacks.

Takeaway 3. *The shadow-free MIAs are more efficient than baseline attacks.*

4.3 Why Shadow-free MIAs Work

In shadow-free MIAs, the attacker needs to calculate and compare two distances for a target user to determine the membership status. To understand why the proposed attacks work, we provide a visualization of $\alpha_1 - \alpha_2$ (see Section 3.3 for the definition of α) for member users and non-member users in Figure 2. In our proposed attacks, 0 essentially is the threshold of $\alpha_1 - \alpha_2$ to determine whether a target user is a member or a non-member. We select four attack settings and calculate the difference value between α_1 and α_2, i.e., $\alpha_1 - \alpha_2$. From the visualization, we can see that the distributions of member and non-member users are very different, explaining why our proposed attacks are highly effective. Specifically, we can observe that in the setting where GRU4Rec is trained on Amazon Beauty, all non-members' data is less than 0, which means that all non-members are correctly classified, which explains that in Table 2, the FPR of the setting is equal to 0. In addition, the dotted line almost overlapped with the mid-line of data distribution of member data in the setting where ICF is trained on MovieLens-1M, which shows that shadow-free MIA can predict about 50% of members correctly for this setting. For the settings where Caser is trained on MovieLens-1M and GRU4Rec is trained on MovieLens-1M, a high portion of members are correctly classified, and less than half of non-members are predicted as members. Due to page limits, we provide the corresponding experimental quantity results in Appendix A.3.

4.4 Ablation Study

We analyse how different parameters of the number of recommends and the length of items' latent feature vectors can influence the attack performance of the proposed attack. Due to page limits, we present the findings from the experiments, while providing the quantity results in Appendix A.3.

The Number of Recommendations n. We use GRU4Rec trained on MovieLens-1M to study how the number of recommendations can influence the attack performance. We vary the number of recommendations from 10 to 100. Experimental results show that increasing the number of recommendations slightly decreases the attack accuracy. This might be because more recommendations can add general popular items to member users' recommendations, making it more difficult to distinguish member users from non-members. Nonetheless, the attack accuracy remains above 97.5% in all cases, demonstrating the high effectiveness of the attack.

The Length of Vectors l. We use GRU4Rec trained on Tafeng to study how the length of the item feature vector can influence the attack performance. We vary the number of length of item feature vectors from 10 to 100. Similar to the study of the number of recommendations, the attack performance is stable at different lengths, and the attack accuracy is all above 98%, demonstrating the high effectiveness of the attack.

Takeaway 4. *The shadow-free MIAs are stable and effective under different attack settings. The number of recommendations and the length of the item feature vector only slightly influence the attack performance.*

5 Discussion

As demonstrated in Section 4, shadow-free MIAs can infer the membership privacy of a user efficiently and effectively. As our attacks only need black-box access to the model, the experimental results in this paper shed light on the vulnerabilities of recommender systems in leaking the membership privacy of their member users. Two possible defenses can be considered to mitigate MIAs. The first one is leveraging differential privacy [Dwork *et al.*, 2006], the most widely used privacy mechanism, to train a differentially private recommender system that should not remember the details of a specific user. For deep learning-based recommender systems, DP-SGD [Abadi *et al.*, 2016] can be leveraged during the training process. Another potential defense is to add randomness to the recommendations for non-members. For example, except for recommending general popular items to non-member users, the recommender system can randomly select some items from the whole recommendation lists and add them to the final recommendations. This can make non-member users' recommendations similar to personalized recommendations of member users, which may make the attacker mistakenly predict non-members as members, reducing the reliability of the MIAs.

Limitations. One limitation of our method is that the target recommender system leverages the popularity-based recommendation strategy to handle the cold-start problem for new users. While this simple yet effective strategy is still a mainstream solution [Sedhain *et al.*, 2014], more strategies employing the information from other sources (e.g., users' social data [Sedhain *et al.*, 2017]) to approximate user preference have also been proposed. We believe that the attack principle in our method can also be applied in such a setting, but some tweaks might be necessary. We leave this part of research in our future work.

6 Conclusion

In this paper, we proposed shadow-free MIAs that can effectively and efficiently infer the membership privacy of a user in recommender systems. Compared to existing works that require training a shadow model on a shadow dataset, our attack requires only black-box access to the target recommender system. We conduct extensive experiments on three benchmark datasets across five recommender systems under different attack settings. The experimental results validate that our attack can efficiently and effectively achieve high attack accuracy at a low false positive rate, which is far better than baseline attacks. The findings in this paper shed light on the vulnerability of recommender systems, emphasizing the importance of comprehensively evaluating the privacy risks of recommender systems. Two possible defenses that can mitigate the membership privacy leakage of recommender systems are discussed. We leave the evaluation of the effectiveness of such defenses in future works.

Acknowledgments

Dr Xuyun Zhang is the recipient of an ARC DECRA (project No. DE210101458) funded by the Australian Government. This work is partially supported by ARC Discovery Project DP200101441. Professor Lianyong Qi is supported by Natural Science Foundation of Shandong Province (No. ZR2023MF007).

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, pages 308–318, 2016.

[Burke, 2002] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-adapted Interaction*, 12:331–370, 2002.

[Carlini *et al.*, 2019] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*, pages 267–284, 2019.

[Carlini *et al.*, 2021] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security*, pages 2633–2650, 2021.

[Carlini *et al.*, 2022] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *S&P*, pages 1897–1914. IEEE, 2022.

[Chen *et al.*, 2017] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*, pages 335–344, 2017.

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Theory of Cryptography Conference (TCC)*, pages 265–284. Springer, 2006.

[Fan *et al.*, 2019] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *WWW*, pages 417–426, 2019.

[Harper and Konstan, 2015] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm Transactions on Interactive Intelligent Systems (TIIS)*, 5(4):1–19, 2015.

[Hayes *et al.*, 2017] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

[He *et al.*, 2017] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *WWW*, pages 173–182, 2017.

[He *et al.*, 2020] Yang He, Shadi Rahimian, Bernt Schiele, and Mario Fritz. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation. In *Computer Vision–ECCV*, pages 519–535. Springer, 2020.

[He *et al.*, 2021] Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. Node-level membership inference attacks against graph neural networks. *arXiv preprint arXiv:2102.05429*, 2021.

[Hidasi *et al.*, 2015] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

[Hu *et al.*, 2022] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[Koren *et al.*, 2009] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[Liu *et al.*, 2018] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. Stamp: short-term attention/memory priority model for session-based recommendation. In *KDD*, pages 1831–1839, 2018.

[McAuley *et al.*, 2015] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, pages 43–52, 2015.

[Narducci *et al.*, 2015] Fedelucio Narducci, Cataldo Musto, Marco Polignano, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. A recommender system for connecting patients to the right doctors in the healthnet social network. In *WWW*, pages 81–82, 2015.

[Pardau, 2018] Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68, 2018.

[Ribeiro *et al.*, 2015] Mauro Ribeiro, Katarina Grolinger, and Miriam AM Capretz. Mlaas: Machine learning as a service. In *ICMLA*, pages 896–902, 2015.

[Rosen, 2011] Jeffrey Rosen. The right to be forgotten. *Stanford Law Review*, 64:88, 2011.

[Salem *et al.*, 2019] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *NDSS Symposium*. Internet Society, 2019.

[Sarwar *et al.*, 2001] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, pages 285–295, 2001.

[Sedhain *et al.*, 2014] Suvash Sedhain, Scott Sanner, Darius Braziunas, Lexing Xie, and Jordan Christensen. Social collaborative filtering for cold-start recommendations. In *RecSys*, pages 345–348, 2014.

[Sedhain *et al.*, 2015] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. Autorec: Autoencoders meet collaborative filtering. In *WWW*, pages 111–112, 2015.

[Sedhain *et al.*, 2017] Suvash Sedhain, Aditya Menon, Scott Sanner, Lexing Xie, and Darius Braziunas. Low-rank linear cold-start recommendation from social data. In *AAAI*, volume 31, 2017.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *S&P*, pages 3–18. IEEE, 2017.

[Song and Mittal, 2021] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security*, pages 2615–2632, 2021.

[Song and Shmatikov, 2019] Congzheng Song and Vitaly Shmatikov. Auditing data provenance in text-generation models. In *KDD*, pages 196–206, 2019.

[Sun *et al.*, 2019] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM*, pages 1441–1450, 2019.

[Tang and Wang, 2018] Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WSDM*, pages 565–573, 2018.

[Tang *et al.*, 2016] Jiliang Tang, Charu Aggarwal, and Huan Liu. Recommendations in signed social networks. In *WWW*, pages 31–40, 2016.

[Wang *et al.*, 2022] Zihan Wang, Na Huang, Fei Sun, Pengjie Ren, Zhumin Chen, Hengliang Luo, Maarten de Rijke, and Zhaochun Ren. Debiasing learning for membership inference attacks against recommender systems. In *KDD*, pages 1959–1968, 2022.

[Wu *et al.*, 2019] Le Wu, Peijie Sun, Yanjie Fu, Richang Hong, Xiting Wang, and Meng Wang. A neural influence diffusion model for social recommendation. In *SIGIR*, pages 235–244, 2019.

[Ye *et al.*, 2022] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *CCS*, pages 3093–3106, 2022.

[Yeom *et al.*, 2018] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *CSF*, pages 268–282. IEEE, 2018.

[Zhang *et al.*, 2019] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

[Zhang *et al.*, 2021] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. Membership inference attacks against recommender systems. In *CCS*, pages 864–879, 2021.

[Zhou *et al.*, 2018] Meizi Zhou, Zhuoye Ding, Jiliang Tang, and Dawei Yin. Micro behaviors: A new perspective in e-commerce recommender systems. In *WSDM*, pages 727–735, 2018.

[Zhou *et al.*, 2019] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *AAAI*, volume 33, pages 5941–5948, 2019.

[Zhu *et al.*, 2023] Zhihao Zhu, Chenwang Wu, Rui Fan, Defu Lian, and Enhong Chen. Membership inference attacks against sequential recommender systems. In *WWW*, pages 1208–1219, 2023.