[1]

Membership Inference Attacks on Recommender Systems: A Survey

JIAJIE HE*, University of Maryland, Baltimore County, USA
XINTONG CHEN*, University of Cincinnati, USA
XINYANG FANG, University of Southern California, USA
MIN-CHUN CHEN, University of Maryland, Baltimore County, USA
YUECHUN GU, University of Maryland, Baltimore County, USA
KEKE CHEN, University of Maryland, Baltimore County, USA

Recommender systems (RecSys) have been widely applied across E-commerce, finance, healthcare, and social media and have become increasingly influential in shaping user behavior and decision-making, underscoring their growing impact across domains. Since RecSys heavily relies on user data, its privacy concerns are significant and need to be addressed urgently. Recent studies on membership inference attacks (MIAs) in RecSys highlight this need. MIAs aim to infer whether a user or an interaction record was used to train a target ReSys model. The success of MIA can lead to severe privacy breaches, e.g., inferring users' special lifestyles. MIAs in RecSys have features distinct from other MIAs in classification models or large language models. However, no systematic survey on this topic has yet been conducted. We present the first comprehensive survey on RecSys MIAs, exploring their taxonomy, design principles, evaluation methods, and defense mechanisms. Based on the summary of existing studies in this area, we also outline several promising future research directions. This survey will raise awareness of privacy risks among RecSys researchers, practitioners, and users, and promote privacy protection practices in RecSys design.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Do, Not, Use, This, Code, Put, the, Correct, Terms, for, Your, Paper

ACM Reference Format:
Jiajie He, Xintong Chen, Xinyang Fang, Min-chun Chen, Yuechun Gu, and Keke Chen. 2018. Membership Inference Attacks on Recommender Systems: A Survey. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 29 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recommendation systems (RecSys) have seen significant advances over the past decade and are widely used across various applications, such as job matching [11], e-commerce [10], entertainment [15], and social media [25]. Besides

[1]Both authors contributed equally to this research.

Authors' Contact Information: Jiajie He, University of Maryland, Baltimore County, Baltimore, USA, jiajieh1@umbc.edu; Xintong Chen, University of Cincinnati, Cincinnati, USA, chen3xt@mail.uc.edu; Xinyang Fang, University of Southern California, Los Angeles, USA; Min-chun Chen, University of Maryland, Baltimore County, Baltimore, USA; Yuechun Gu, University of Maryland, Baltimore County, Baltimore, USA; Keke Chen, University of Maryland, Baltimore County, Baltimore, USA.

Manuscript submitted to ACM

advanced algorithm design and powerful computational resources, the availability of large datasets is another key factor contributing to the success of RecSys [90]. As RecSys datasets often contain rich and highly sensitive personal information, such as users' purchase histories, browsing behaviors, watched movies or shows, search queries, clicked items, social connections, demographic attributes (e.g., age, gender, location), and even implicit behavioral patterns (e.g., temporal activity or preference shifts), RecSys model owners must ensure such privacy-sensitive information is not inadvertently leaked through model parameters, intermediate representations, or generated recommendations. However, recent studies [12, 22, 24, 79, 82, 84, 91, 92] have shown that RecSys models are prone to memorizing information of user data, making them vulnerable to several privacy attacks [7, 8, 17, 18, 21, 30, 32, 64, 86]. Among these attacks, **membership inference attacks (MIAs)** are considered as the fundamental step to breach privacy, which aim to infer whether a specific user or an interaction is included in the training data of a RecSys model.

The first MIA on machine learning was proposed by Shokri et al. [64] on several classification models that demonstrated that an attacker can determine whether a data record was used to train an ML model solely from the prediction vector of that record (i.e., with black-box access to the target model). Since then, a growing number of studies have investigated MIAs across various domains, including computer vision [7], natural language processing [49, 58, 80], and audio processing [63]. The research on MIAs in RecSys has started relatively late, compared to other domains. However, since its impacts are more widespread, and RecSys MIA attacks have unique features distinct from other MIA attacks, there is an urgent need to understand them and design mitigation methods.

Unique features. We list several unique features of RecSys MIAs as follows.

- The nature of RecSys leads to diverse MIA targets at different levels. So far, researchers have discussed the attacks targeting the user level, the interaction level, and the social level, i.e., the connection between users. These dimensions are not seen in other types of MIAs.
- Adversarial knowledge is different. Traditional MIA techniques rely on posterior probabilities, which are often unavailable in recommendation settings. In practice, adversaries can only observe the ranked lists of items produced by the RecSys, rather than the underlying prediction scores or confidence values.
- New attack vectors and system parameters. RecSys may use unique global information, such as user and item embeddings. Its output setting, i.e., the number of recommended items, also introduces an additional layer of complexity for attack design.
- RecSys has many different designs, such as matrix-factorization-based, graph-based, sequence-based, and federated RecSys. Each may impose unique challenges to MIAs, requiring particular attack designs.

Direct impacts on individuals. A privacy breach in RecSys also has great impacts on individuals due to the widespread deployment and the large user base compared to other systems. For example, identifying that specific purchase records were used to train an e-commerce RecSys may expose a user's preferences or behavioral traits. Exposed medicine recommendations may reveal sensitive medical conditions, such as HIV or syphilis, causing significant social or psychological harm. The National Institute of Standards and Technology (NIST) formally classifies MIAs that reveal an individual's presence in a training dataset as a privacy and confidentiality violation [70], and such risks place substantial regulatory pressure on RecSys providers under laws including GDPR [76], CCPA [1], and PIPL [2]. Recent real-world incidents further underscore the severity of these threats: in 2023, a Spotify API exposure allowed unauthorized access to users' private playlists and listening histories, enabling unintended profiling, while in 2024, researchers showed that TikTok's "For You" algorithm could leak sensitive attributes, such as location, interests, and social ties, through latent

embeddings. Together, these cases illustrate that even well-engineered recommendation systems may inadvertently disclose personal information, undermining user trust and highlighting the urgent need for robust privacy defenses.

We present the first systematic survey that comprehensively summarizes existing membership inference attacks and defense mechanisms in recommendation systems. Specifically, we establish a taxonomy of MIA approaches across multiple dimensions and analyze their theoretical foundations, methodologies, evaluation protocols, emerging challenges, and future research directions to guide the development of privacy-preserving RecSys. A closely related survey, Hu et al. [31] in 2022, describes the MIAs in general, covering only one paper in RecSys MIA. Since then, more unique features of RecSys MIAs have been explored, and thus, they deserve a dedicated, more systematic in-depth analysis. Our extensive, up-to-date literature search and analysis are timely and address this urgent need. The main contributions of this article are summarized as follows:

- **Comprehensive Review.** To the best of our knowledge, this is the first work to provide a comprehensive review of membership inference attacks and related defenses on RecSys models. In this work, we establish novel taxonomies of membership inference attacks and defenses, respectively, based on various criteria.
- **Datasets and Metrics.** We summarize the evaluation resources for MIAs on RecSys, including the commonly used datasets, recommendation models, and evaluation metrics regarding each design principle. By providing a clear mapping of these resources, we aim to help researchers select the appropriate tools to evaluate the effectiveness of different MIA approaches.
- **Challenges and Future Direction.** MIAs on RecSys is an active and ongoing area of research. Based on the literature reviewed, we have discussed the challenges yet to be solved and proposed several promising future directions for MIAs designed on RecSys to inspire interested readers to explore this field in more depth.
- **Online Updating Resource.** We create an open-source repository[1] that includes most, if not all, the relevant work. This repository provides links to all papers and released code to help researchers interested in this area. As a small number of the surveyed papers are only available as preprints, authors are welcome to update us when the full publication information becomes available. We will continue to update the repository with new work in this domain. We hope this open-source repository will shed light on future research on membership inference analysis in RecSys.

The rest of the article is organized as follows: Section 2 introduces MIAs on RecSys preliminaries. Section 3 introduces the existing attack approaches and provides taxonomies to categorize the released papers. In Section 4, we discuss current defenses on RecSys MIAs. Section 6 discusses the challenges and proposes future directions. Section 7 concludes this article.

2 Preliminaries

2.1 Recommendation System (RecSys)

Recommendation systems (RecSys) have undergone remarkable development over the past decade and have been extensively deployed across a wide range of application domains, including job matching [11], e-commerce [10], and online entertainment [15]. By analyzing and modeling complex user–item interaction patterns [25, 26, 61], RecSys can accurately predict user preferences and deliver personalized recommendations, thereby enhancing user satisfaction, engagement, and platform profitability.

[1] https://github.com/Richardwarriors/Membership-Inference-Attacks-on-Recommendation-System

Early RecSys primarily relied on shallow models such as matrix factorization (MF), which represented users and items in low-dimensional latent spaces to capture collaborative signals. However, as the scale and diversity of data grew, traditional shallow models struggled to capture the nonlinear and heterogeneous nature of user behavior. This limitation ushered in the deep learning era of RecSys, leading to the emergence of models such as NeuMF [26], LightGCN [25], and SimpleX [50], which leverage neural architectures and graph structures to better represent complex user–item relationships.

Recently, with the advent of large language models (LLMs), a new paradigm of LLM-based RecSys has emerged. Representative frameworks such as P5 [16], M6-Rec [13], and TALLRec [5] integrate the powerful natural language understanding and generative capabilities of LLMs to enhance recommendation quality, interpretability, and user interaction. These models signal a shift from feature-based to instruction- and context-driven recommendation generation.

However, as RecSys becomes increasingly large-scale, data-hungry, and ubiquitous in real-world applications, ensuring its reliability, fairness, and privacy has become a critical research priority. User data in recommendation systems often contains sensitive, personally identifiable information (e.g., purchase histories, viewing behaviors, social links, and demographic traits), making these systems vulnerable to various privacy attacks. Among them, membership inference attacks (MIAs) have received growing attention, as they enable adversaries to determine whether a user's data was included in a model's training set—posing serious risks to user privacy and organizational trust.

In the following section, we present a detailed and systematic discussion of MIAs in RecSys, outlining their principles, attack models, defense mechanisms, and open research challenges.

2.2 Membership inference attacks (MIAs).

Given a trained target model f and a target record z, an adversary aims to determine whether z was included in the training dataset of f. This can be formulated as a binary hypothesis testing problem:

$$H_0 : \ z \notin \mathcal{D} \quad \text{vs.} \quad H_1 : \ z \in \mathcal{D},$$

where $\mathcal{D}$ denotes the training set. The adversary computes a decision statistic based on information obtained from the target model—such as the output confidence[65], likelihood ratio[7], or training loss trajectory[38]—and compares it against a threshold to decide between the two hypotheses. Intuitively, records that the model has previously seen (members) tend to produce different output characteristics than unseen records (non-members), due to phenomena such as model overfitting or memorization. Therefore, a membership inference attack (MIA) can be interpreted as a model-based distinguishing attack that exploits these behavioral discrepancies to infer the presence of specific records in the training data.

3 Membership Inference Attacks on Recommendation Systems

In this section, we first give a general definition of MIAs on RecSys and then introduce adversarial knowledge, attack approaches, and target models. We will further explain in detail how the unique features of RecSys play in RecSys MIAs.

3.1 Definition of MIAs on RecSys

To better illustrate the definition of MIAs on RecSys, we introduce a typical framework of MIAs on RecSys, shown in Figure 1. The attacker uses the designed MIA methods to attack the trained RecSys and the definition of MIAs on RecSys is as follows: Given an exact input user information, an attacker infers whether the user information is used in training data.

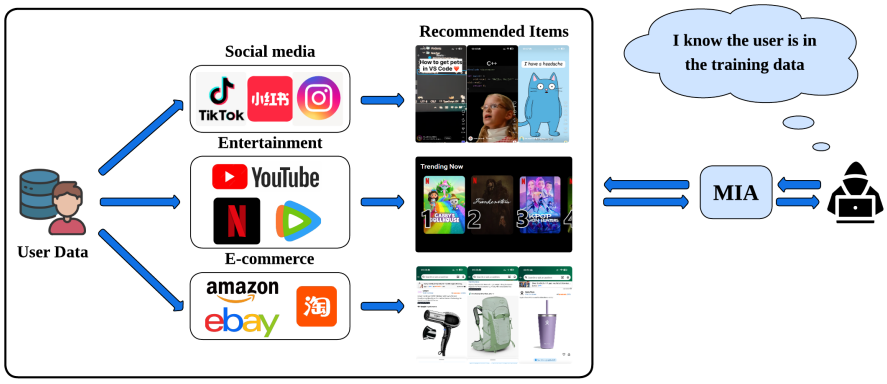

Fig. 1. The Framework of MIA in Recommendation System.

3.2 Threat Models

The amount and type of information available to an attacker critically determine the feasibility and strength of membership inference attacks (MIAs) against recommendation systems (RecSys). In this section, we first formalize the attacker's knowledge and then describe black-box and white-box MIAs in the RecSys setting. Broadly speaking, two classes of background knowledge are most relevant: (i) knowledge about the training data distribution, and (ii) knowledge about the target model.

Knowledge about the training data distribution. It refers to the attacker's understanding of the distribution from which the model was trained. Many MIA formulations assume that the adversary can access or synthesize a shadow dataset drawn from the same distribution as the target training data. This assumption is justified in practice: when distributional statistics are available, the shadow dataset can be generated via statistics-based synthesis, whereas in other cases, one may employ model-based synthesis techniques to approximate the underlying distribution [64]. For nontrivial evaluation, it is typically assumed that the shadow dataset is disjoint from the target training set.

Knowledge about the target model It captures information about how the RecSys is trained and parameterized, including posterior probability, and hyper-parameters (e.g., negative-sampling ratio).

RecSys raises an additional practical requirement not commonly emphasized in other domains: besides a shadow dataset, many RecSys attacks assume access to an *item-embedding generating* (IEG) dataset that is disjoint from both the target and shadow datasets. The IEG dataset is used to produce item embeddings for the full item catalog, a step necessary because items are typically known a priori, even when user interaction traces are private. Requiring a separate IEG dataset is a mild and realistic assumption—platforms and public catalogs make item descriptions readily available—yet it materially affects attack design and transferability. Based on adversarial knowledge, we can characterize the dangerous levels of existing attacks.

White-box Attacks. Under this setting, an attacker can get some inaccessible information and use it to attack a target RecSys model. The information includes the posterior probability of the item and the learned parameters of the target model.

Black-box s. In this case, an attacker can only have black-box access to a target model. The attacker is given information limited to training data distribution, the user-interacted item set, and the recommended item set.

Figure 2 illustrates the conceptual distinction between white-box and black-box MIAs on a target RecSys model. A green tick (✓) indicates available information, while a red cross (✗) denotes inaccessible information. As shown, the white-box adversary possesses access to detailed model internals, including learned parameters and posterior probabilities, whereas the black-box adversary is restricted to external query responses and user interaction data.

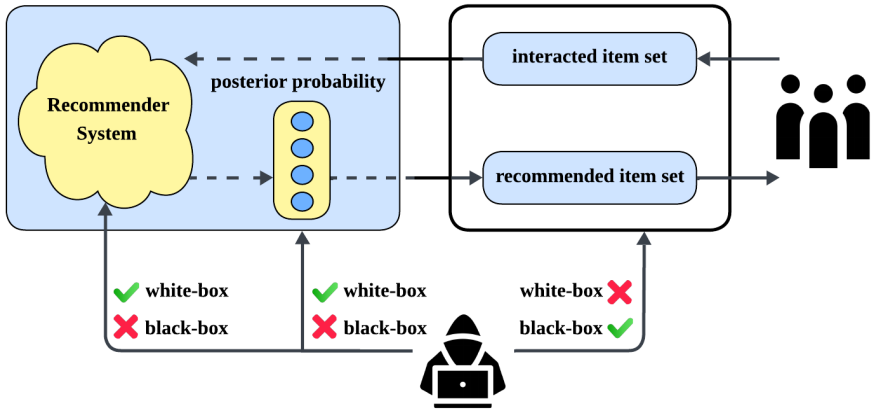

Fig. 2. The Overview of white-box and black-box MIA in Recommendation System.

In the real world, obtaining access to model parameters or training configurations is exceedingly rare. Consequently, most recent research on MIAs in RecSys focuses on the black-box setting, which better reflects realistic deployment scenarios. The white-box setting, by contrast, is primarily explored in the context of Federated RecSys [82], where parameter updates may be partially exposed to participating clients. In traditional centralized RecSys models, white-box analyses are often used from a developer's standpoint to assess privacy risk scoring [22] and proactively safeguard users identified as privacy-sensitive.

Although black-box attackers operate with significantly less information, the fact that effective black-box MIAs can still succeed underscores a critical vulnerability in modern RecSys. Demonstrating attack success under such limited information further emphasizes the urgent need for robust privacy-preserving mechanisms in RecSys.

3.3 Taxonomies of Membership Inference Attacks on Recommendation Systems

To give readers a general picture of MIAs and help readers find the most relevant papers easily, we create a taxonomy of MIAs on RecSys in Figure 3. In this taxonomy, we categorize all released MIA papers by attack strategy and target model. Specifically, for papers in the target model level, we further categorize them by target ML model type, e.g., Matrix-factorization-based RecSys, Sequential RecSys, Graph-based RecSys, LLM-based RecSys, etc. For papers in the attack strategies level, we further categorize them by specific attack levels, i.e., user level, interaction level, and social level. For papers in the adversarial knowledge category, we further divide them into black-box and white-box attacks. Lastly, for papers in the algorithmic level category, we further divide them based on whether the target models are trained in a centralized or federated manner. Note that Figure 3 not only gives general taxonomies for MIAs according to the above criteria, but also provides detailed characteristics for specific categorized papers.

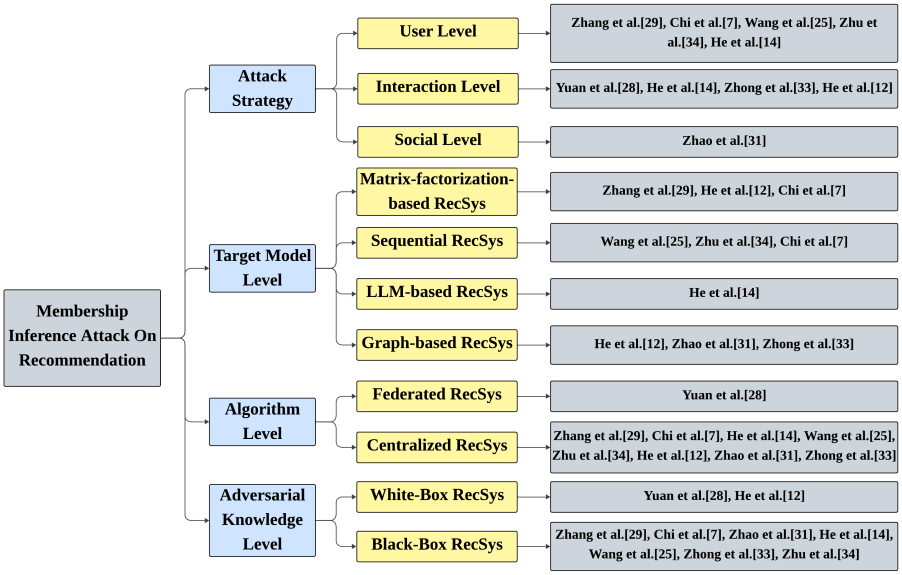

Fig. 3. Taxonomy: Membership Inference Attack on Recommendation System.

3.4 Targets of Membership Inference Attacks on Recommendation Systems

Recommendation systems are inherently complex information-fusion systems that integrate diverse sources of user and item data. As a result, they encapsulate multiple levels of privacy-sensitive information from the adversary's perspective. The training dataset for a RecSys typically contains rich, heterogeneous information, including user attributes, behavioral histories, and social relationships. From these different aspects, adversaries can launch various membership inference attacks (MIAs) targeting distinct forms of private information. From the user level, the adversary observes the behavior of the target RecSys on member records (i.e., data points used during training) versus non-member records (i.e., data points unseen during training) to infer whether a specific user was included in the training dataset. From the interaction level, the focus shifts to user behaviors and preferences, which tend to be more sensitive. Here, the attacker aims to infer whether a particular user–item interaction (e.g., a purchase, click, or rating) was part of the training data, thereby revealing a user's interests, preferences, or even daily habits. Finally, at the social level, which arises in social or graph-based recommendation systems, the adversary attempts to exploit the collaborative filtering or graph embeddings associated with recommended items to infer hidden social ties—such as whether two users are connected in the underlying social network.

Overall, these three attack granularities—**user-level**, **interaction-level**, and **social-level** MIAs—represent the principal dimensions along which privacy risks manifest in modern RecSys.

3.4.1 User-Level MIAs. Zhang et al.[84] were the first to propose the user-level membership inference attack in RecSys based on item embedding differences, aiming to infer whether a user's data was included in the training dataset of the RecSys model. Under the black-box setting, the attacker can only observe the output recommendation list. The item-embedding-based method assumes that if a user's data is included in the training set, the recommended items should be similar to the items the user has interacted with. In contrast, for non-member users, the system lacks

knowledge of their preferences and therefore cannot generate personalized recommendations. In the original work, the top popular items were used as the recommendations for non-members. To construct attack features, the embeddings of the interacted items were averaged, and those of the recommended items were averaged; the difference between the two was then used as the membership feature. A two-layer multilayer perceptron (MLP) was trained as the attack model to distinguish members from non-members. However, this approach is unstable when the target and shadow datasets differ. The reason is that the method implicitly assumes that item embeddings are generated by a fixed collaborative filtering model. In practice, embeddings trained by different algorithms (e.g., BPR vs. NeuMF) can diverge in the latent space even when trained on the same dataset. Furthermore, the item embeddings generated by the target and shadow models may also exhibit discrepancies, which degrade attack performance. To address this issue, Wang et al. [79] proposed a debiased learning MIA (DL-MIA) framework for RecSys that mitigates the embedding bias between the target and shadow models. For DL-MIA, to overcome the limitations of Item Difference MIAs (ID-MIA), which suffer from (i) a training-data bias—distributional gaps between shadow and target recommenders that make attack samples generated by the shadow model poorly transferable—and (ii) an estimation bias—since the attacker cannot observe hidden states (user/item embeddings), externally constructed difference vectors are noisy—the DL-MIA framework explicitly debiases learning within the difference-vector paradigm. Concretely, it (1) builds difference vectors from each user's history and the system's recommendation list; (2) employs a variational auto-encoder–based disentangled encoder to separate recommender-invariant from recommender-specific features, narrowing the shadow–target gap and mitigating training-data bias; (3) learns a truth-level score per difference vector as a sample weight to discount poorly estimated features, thereby mitigating estimation bias; and (4) trains a member/non-member classifier (MLP) on the disentangled, reweighted representations, optimized with an alternating training procedure. Empirically, DL-MIA simultaneously reduces both biases and achieves state-of-the-art attack performance across general and sequential RecSys.

Although subsequent experiments demonstrated that DL-MIA is more accurate and stable than the ID-MIA under target–shadow mismatches, its effectiveness remains sensitive to the information exposed to the adversary—most notably the dimensionality of item embeddings and the length of the top $-K$ recommendation list. Empirically, performance tends to improve with larger embedding dimensions and larger K, whereas short lists (e.g., $K \leq 10$) often provide a weak signal in practice. Moreover, obtaining the training dataset distribution is unrealistic. Chi et al[12]. proposed the shadow-free MIA method to infer the membership without training the shadow model, compared with the previous methods[79, 84, 92]. The intuition of the SF-MIA is to compare the recommended items with general popular items (which can be obtained by creating an empty account). If the recommended items align more closely with popular items, then it's likely not a member. But if the recommended items have a higher similarity to the user's historical interactions, then it'll be classified as a member. The experiment results show that this shadow-free approach not only significantly reduces computational cost for MIAs but also reaches a comparable performance to previous shadow-training approaches. At the same time, since the shadow model does not need to be trained, this method reduces the attack's time cost. However, the common assumption that non–member recommendations align with a globally popular item set is brittle in the real world. The current sequential RecSys can provide personalized recommendations for both member and non-member users. To address the member/non-member mode gap, Zhu et al.[92] propose the Model Extraction based MIA (ME-MIA) for sequential recommenders, which first extracts a surrogate model via black–box queries that align its recommended item set and rank order with the target using generic list–consistency objectives, and then performs membership inference using the surrogate's richer signals (scores, ranks, similarities). ME-MIA targets sequential recommenders in a black-box setting by first extracting a surrogate model that imitates

the target's top $-K$ items and their rank order using two generic objectives—ranking consistency and positive-item consistency—derived solely from recommendation lists. Because the surrogate's training does not reveal membership labels, a shadow model (trained with the same objective) is used to construct labeled data; rich signals from the surrogate (scores, ranks, similarities) then feed a binary classifier for membership inference. To reduce data demands, ME-MIA offers (i) a data-efficient variant that augments sequences by replacing actions with nearest-neighbor items in the surrogate's embedding space, and (ii) a data-free variant that synthesizes sequences autoregressively and queries the target for soft labels, following data-free model-extraction practice. These steps yield effective, transferable attacks in black-box sequential settings. While ME-MIA demonstrates that membership inference is possible even in data-free settings, this comes at the cost of effectiveness; moreover, the role of item scoring in shaping attack success remains underexplored, and it is unclear whether ME-MIA readily generalizes to non-sequential RecSys. Despite recent progress at the user level, open challenges persist: how to design more efficient attacks, reduce confounding factors that obscure the signal (e.g., embedding dimensionality, top $-K$, popularity effects), and accommodate the inherent heterogeneity of recommender systems with model–agnostic methods. Beyond traditional RecSys MIAs, He et al.[24] (to our knowledge) present the first user–level attacks for *LLM–based* recommender systems, proposing five attacks—*Inquiry*, *Hallucination*, *Semantic*, *Poisoning*, and *Contrast*—that exploit LLMs' memorization, text generation, and hallucination behaviors; evaluated across multiple popular LLMs, these attacks demonstrate clear effectiveness in revealing whether specific user were included in in–context prompts.

3.4.2 Interaction-level MIAs. RecSys poses privacy risks beyond the user level: in addition to concerns about whether a user is included in the training set, one may also be interested in the presence or absence of specific user–item interactions, referred to as *interaction-level privacy*. In this part, we focus on centralized RecSys, while the discussion of Federated RecSys will be presented in Section 3.6. Compared to user-level privacy studies, research on interaction-level privacy remains relatively limited. Although the privacy risks associated with interaction-level breaches are often more severe, designing effective MIAs at this level is considerably more challenging, primarily because existing embedding-difference-based methods are not directly applicable. To address the challenge of interaction-level MIA on RecSys, Zhong et al.[91] firstly proposed interaction-level MIA called MINER on a knowledge-graph (KG)–based RecSys, which is a framework that learns to infer whether a specific user-item interaction was included in the training data by leveraging knowledge-enhanced embeddings and a bilateral-branch attack model. The intuition of MINER is measuring the distance similarity metric between the interacted item and recommended item to infer the member and non-member data. For ranked recommendation lists, MINER computes a discounted similarity score (DS) that logarithmically weights higher-ranked items more heavily:

$$DS(i, i') = \frac{d(\mathbf{e}_i, \mathbf{e}_{i'})}{\log_2(r_{i'} + 1)},$$

where $d(\cdot)$ denotes a distance function and $r_{i'}$ is the ranking of item i' in the top-k recommendation list. Multiple distance metrics (L1, L2, cosine, Bray–Curtis) are used, and their concatenation forms the feature vector $\mathbf{x}$ for each user–item pair.

Considering the distribution of the item, the author thinks the personalized interacted item (non-popular item) contains more sensitive information. For example, consider a healthcare RecSys. Determining whether a patient has been treated ("interacted") with HIV (a rare disease) carries greater sensitivity than discerning whether the user has been treated with flu, a common disease. MINER introduces a bilateral-branch attack model with two sub-networks: a *main branch* trained on the original long-tailed distribution and a *regularizer branch* trained on a re-balanced distribution,

which enables each branch to learn complementary knowledge from head and tail interactions, respectively. Through this bilateral learning strategy, MINER effectively mitigates the influence of long-tailed distributions and achieves high attack accuracy on both head and tail interactions.

However, while MINER targets the long-tail, it shows limited effectiveness in the low-FPR regime and its reliance on KGs constrains applicability to other RecSys architectures. To address the model-agnostic constraint, He et al. [22] adapt the Likelihood Ratio Attack (LiRA) [7] to the recommender systems setting and propose *RecLiRA*. The core idea is that the posterior confidence distributions of member (IN) and non-member (OUT) samples exhibit measurable differences, enabling privacy risk to be quantified through their statistical separability. For each shadow model, RecLiRA collects confidence scores for IN and OUT samples and models them as Gaussian distributions. A statistic $q = |2p - 1|$ and its logit transformation $\phi(q) = \log!\left(\frac{q}{1-q}\right)$ are then used to better distinguish between the two distributions. RecLiRA is versatile and can be applied to both *interaction-level* and *user-level* membership inference attacks.

In addition to traditional RecSys, the development of LLM-based RecSys in recent years has also led people to pay attention to the privacy issues associated with LLM-based RecSys. He et al.[24] pioneer interaction–level MIAs for LLM-based RecSys, exploiting LLMs' tendency to memorize prompt content through *direct inquiry* and *contrast* attacks, underscoring practical risks for In-Context Learning RecSys.

3.4.3 Social-Level MIAs. Unlike user-level and interaction-level MIAs, which only require knowing whether a user or a user's interactions was included in the training dataset, Zhao et al.[89] introduces a social-level Membership Inference Attack (SMIA) framework, which moves beyond traditional user- and interaction-level attacks to infer whether a social link exists between two users in a social RecSys. The framework targets the inference of whether a user pair (u_1, u_2) has a social relation in the social graph G_S of a social RecSys. The intuition is that users with social ties tend to have higher similarity in recommendation lists or embedding space (i.e., social homophily), allowing inference of hidden relationships. The attacker has black-box access to the target RecSys model M_{target} (i.e., only recommendation results), and optionally a shadow social graph G'_S. The adversary firstly collects the recommendation outputs (top-k lists) from M_{target} for target users and forms a *shadow interaction graph* G'_A. This graph approximates how users are connected via item recommendations. Then, using G'_S and G'_A, the attacker trains a dual-branch model:

- *Shadow Social Preference Learner:* takes the social graph G'_S and computes user representations via a GCN, aiming to capture social influence and homophily patterns (embeddings E^S).
- *Shadow Behavioral Preference Learner:* uses the interaction graph G'_A to extract user behaviors independently (embeddings E^B).

The embeddings from both branches are then aggregated (e.g., concatenation, attention) to form a combined pair-feature representation for any user pair. Finally, the attacker train the binary classifier to predict the $\hat{y}_{u_1,u_2}$. If a social relation exists between u_1 and u_2, $\hat{y}_{u_1,u_2} = 1$, else 0. In short, SMIA reveals a previously underexplored privacy dimension social-level inference in RecSys—by combining recommendation output analysis, shadow-model learning, and user-pair classification to infer hidden social ties.

3.5 Membership Inference Attacks on different Recommendation System

3.5.1 Matrix-Factorization based RecSys. Matrix factorization–based RecSys (MF-RecSys) represent users and items as low-dimensional latent vectors and predict user preferences through simple interactions (e.g., inner products) that capture the co-occurrence structure within the user–item matrix. MF-based approaches have been widely adopted

across domains such as movie recommendation, e-commerce, music, and news, and remain strong baselines for both explicit-rating and implicit-feedback settings (e.g., BPR for pairwise ranking [61], LFM, and NeuMF [26]).

Zhang et al. [84] introduced the earliest membership inference attacks (MIAs) in the RecSys domain at the user level, which directly apply to MF-style models. To mitigate distributional bias between the target dataset and shadow dataset, Wang et al. [79] proposed a debiased learning MIA (DL-MIA) framework that reduces the discrepancy in item embeddings generated by different methods. Although these MIAs demonstrated strong attack performance, they depend on the construction of shadow or surrogate models, thereby increasing both the attack complexity and computational overhead. To overcome these limitations, Chi et al. [12] proposed a shadow-free MIA (SF-MIA) that infers membership directly without training a shadow model, achieving comparable performance to prior approaches. The underlying intuition of these MF-based MIAs lies in analyzing the embedding differences between the interacted item set and the recommended item set for member and non-member data. Building upon this direction, He et al. [22] designed a LiRA-based interaction-level attack (RecLiRA) that achieves state-of-the-art true positive rates (TPR) at low false positive rates (FPR) across common RecSys architectures. Moreover, they introduced a differential-privacy–inspired privacy score, ln(TPR/FPR), to quantify interaction-level risk and aggregate it to the user level [23]. The intuition behind RecLiRA is to leverage the posterior probability confidence differences between member and non-member data to assess and exploit privacy vulnerability.

3.5.2 MIA on Sequential-based RecSys. Sequential-based RecSys aims to model users' dynamic preferences by capturing the temporal dependencies and ordering of their historical interactions. Unlike traditional collaborative filtering methods that treat user–item interactions as unordered sets, sequential models leverage ordered interaction sequences to predict the next item a user is likely to engage with. Sequential-based RecSys have been successfully applied in various domains such as e-commerce [36], music streaming [27], and online content recommendation [73]. Early approaches employed recurrent neural networks (RNNs) and gated recurrent units (GRUs) to model sequential patterns [28], while more recent methods adopt self-attention mechanisms, such as SASRec [36] and BERT4Rec [68], to capture long-range dependencies and contextual relationships between user actions. Similar to other deep neural architectures, sequential models may inadvertently memorize sensitive user behavior, making them vulnerable to membership inference attacks (MIAs) that exploit temporal behavioral discrepancies between training and non-training sequences. Wang et al. [79] proposed a debiased learning MIA (DL-MIA) framework, which performs user-level membership inference by analyzing embedding differences between members and non-members. Considering that attackers in real-world scenarios rarely have access to the true distribution of training data, Zhu et al. [92] introduced the Model Extraction–based MIA (ME-MIA) for sequential RecSys, which operates in a black-box setting by first extracting a surrogate model that replicates the target's ranking behavior and then utilizing the surrogate's rich signals (e.g., scores, ranks, similarities) to train a binary classifier for membership inference. ME-MIA further proposes data-efficient and data-free variants to reduce reliance on real user sequences while maintaining high attack effectiveness and transferability.

Although these MIAs have achieved strong performance, they require additional shadow or surrogate models, increasing attack complexity and computational overhead. To address this limitation, Chi et al. [12] proposed a shadow-free MIA (SF-MIA) that infers membership without training a shadow model. SF-MIA determines membership by comparing a user's recommended items with general popular items—closer alignment indicates non-membership, whereas higher similarity to the user's historical interactions suggests membership. This shadow-free approach significantly reduces computational cost while achieving performance comparable to shadow-model–based attacks. In summary, the success of MIAs on sequential RecSys largely stems from the distinguishable embedding differences

between the interacted item set and the recommended item set for members versus non-members, revealing the inherent privacy vulnerability of temporal modeling in Sequential-based RecSys.

3.5.3 MIA on Graph-based RecSys. Graph-based RecSys model user–item interactions as graphs, where nodes represent users or items and edges correspond to interactions such as ratings, clicks, or purchases. By leveraging the rich relational structure inherent in these graphs, such models aim to learn high-quality embeddings that capture both connectivity patterns and higher-order collaborative signals. Graph-based RecSys have achieved remarkable success in modeling user preferences and item similarities through message-passing and neighborhood-aggregation mechanisms, as demonstrated in representative models such as LightGCN [25], CKE [83], and KGAT [78]. These approaches extend traditional collaborative filtering by propagating information along user–item bipartite graphs to capture multi-hop dependencies.

Despite their effectiveness, the graph-based paradigm introduces new privacy risks. The learned embeddings inherently encode users' interaction behaviors and neighborhood relationships, which can be exploited by membership inference attacks (MIAs). Yuan et al. [82] first proposed an *interaction-level* MIA on Federated RecSys, while Zhong et al. [91] extended this idea to centralized RecSys. However, both of these studies focus solely on the interaction level. More recently, Zhao et al. [89] discovered that edge connections between users can introduce a new form of sensitive information leakage, termed *social-level privacy risk*. In this setting, the adversary's objective shifts from identifying individual user–item interactions to inferring social relationships among users—such as friendships, follow links, or communication patterns—on social media platforms. The intuition behind social-level MIAs lies in the distinguishable collaborative filtering patterns between members and non-members within the social graph, revealing privacy vulnerabilities beyond traditional interaction-level attacks.

3.5.4 MIA on LLM-based RecSys. He et al. [24] introduced the first study of membership inference attacks (MIAs) on large language model (LLM)-based RecSys. Their work focuses on In-Context Learning (ICL)-based RecSys, which are widely adopted in conversational recommendation scenarios to address the user cold-start problem. The authors designed five attack strategies that exploit the generalization, memorization, and reasoning capabilities of LLMs: *direct inquiry*, *contrast*, *semantic*, *hallucination*, and *poisoning* attacks. Among these, the direct inquiry and contrast methods can be applied to both user-level and interaction-level privacy settings. The attack intuition is the difference in memorization degree between members and non-members on LLM.

3.6 Membership Inference Attacks against Federated Learning Recommendation System

Federated learning (FL) has recently emerged as an alternative to conventional centralized learning, where all training data are pooled and a machine learning (ML) model is trained on this joint dataset. FL allows multiple parties to collaboratively train an ML model in an interactive manner without directly sharing their raw data. It is an attractive framework for training models on decentralized and privacy-sensitive data [52, 53]. However, the success of membership inference attacks (MIAs) against FL has shown that FL may still reveal sensitive information and does not always provide sufficient privacy guarantees. Melis et al. [54] introduced the first MIA against FL. Their study focused on a text classification task, where the target models were recurrent neural networks equipped with a word-embedding layer to transform inputs into low-dimensional vector representations through an embedding matrix. The embedding matrix was treated as a parameter of the global model and collaboratively optimized. During training, the gradient of the embedding layer is sparse with respect to the input words; for a given batch of text, the embedding is updated only for the words that appear in the batch, while the gradients of all other words remain zero. The attacker can thus observe

non-zero gradients to infer which words occur in the training data. Although MIAs have been widely investigated and achieved remarkable success in federated classification tasks [54], the existing attack and defense approaches cannot be directly applied to federated recommender systems (Federated RecSys) due to the significant architectural differences between federated classification and Federated RecSys.

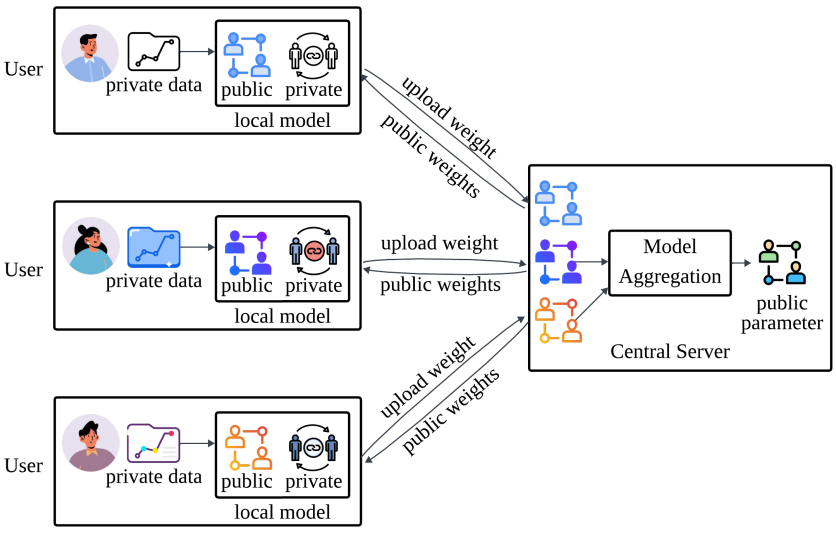

Fig. 4. The Overview of Federated Recommendation System.

For better understanding, the framework of a typical Federated RecSys is illustrated in Figure **??**. Ammad et al. [75] proposed the first Federated RecSys framework based on collaborative filtering, which has inspired many subsequent studies. For example, FedFast [55] aimed to accelerate the convergence of Federated RecSys training, while Imran et al. [33] and Wang et al. [77] focused on improving the efficiency of Federated RecSys. With the rapid progress achieved in a short period, a few recent studies have begun to examine whether Federated RecSys are indeed "safe." For instance, [87] was the first work to analyze the privacy issues in Federated RecSys; it mainly discussed the leakage of sensitive attribute information and proposed an effective protection approach. Although several works [44–46] have studied user information leakage and corresponding defenses in Federated RecSys, the first and only study of MIAs on Federated RecSys was conducted by Yuan et al. [82].

The main challenges of designing MIAs on Federated RecSys arise from two aspects. First, regarding the attack objective, MIAs in federated classification aim to determine whether a given sample has been used in the federated training process and which client has used it for local training. However, in Federated RecSys, the set of items associated with each client can be easily inferred by checking which item embeddings are updated by that client. Nevertheless, this information alone is not meaningful, since the item set contains both positive and negative samples (i.e., interacted and non-interacted items), and only positive samples reveal users' private preferences. Second, from the attack implementation perspective, MIAs in federated classification often rely on extra i.i.d. data, which is infeasible in Federated RecSys. Moreover, the architecture of Federated RecSys is fundamentally different from that of federated classification models: each client in Federated RecSys maintains private parameters (i.e., user embeddings), whereas in federated classification all model parameters are shared among clients.

Yuan et al. [82] conducted the first systematic study of white-box interaction-level membership inference attack on Federated RecSys (IFed-MIA) where a curious-but-honest central server attempts to infer a user's private interaction set $\mathcal{V}_i^+$. The server is assumed to have access only to the public parameters $\mathbf{V}_i^t$ uploaded by each client and basic hyperparameters (e.g., learning rate and negative sampling ratio), without direct access to user embeddings or local data. By analyzing which item embeddings are updated during local training, the server can infer which items a user has interacted with, but cannot distinguish whether those interactions are positive or negative. To infer the actual label r_{ij} of each interaction, the attacker leverages an empirical distance principle: given three locally trained models—M_i on the true dataset $\mathcal{D}_i$, M_i' on $\mathcal{D}_i$ with different initialization, and M_i'' on a reversed dataset $\mathcal{D}_i^j$ where r_{ij} is flipped—it consistently holds that $\text{dist}(\mathbf{v}_j, \mathbf{v}_j') < \text{dist}(\mathbf{v}_j, \mathbf{v}_j'')$. Hence, comparing embedding distances enables the attacker to infer whether an item is positively rated.

Since the server does not know the true ratings in $\mathcal{D}_i$, Yuan et al. constructed a synthetic dataset $\mathcal{D}_i^{\text{fake}}$ by randomly assigning ratings to updated items according to the known negative sampling ratio (e.g., 1:4). The attacker then trains a fake local model M_i^{fake} on $\mathcal{D}_i^{\text{fake}}$ and compares item-embedding distances between $\mathbf{V}_i^t$ and $\mathbf{V}_i^{\text{fake}}$. Items with the smallest distances are labeled as positive, and the process iterates until the target ratio of positive samples is met. The entire inference procedure can be executed asynchronously on the server side without interrupting the standard federated training process. Experimental results on Fed-NCF and Fed-LightGCN show that IMIA achieves over 90% accuracy in predicting user–item interactions across multiple datasets, revealing that even without access to private embeddings or raw data, Federated RecSys remain highly vulnerable to fine-grained interaction-level privacy leakage. However, training hyper-parameters such as the negative sampling ratio are typically not observable to an external adversary in practice, it remains to be seen whether IFed-MIA can be effectively used in real-world offenses.

4 Defense Mechanisms

Although the research on membership inference attacks (MIAs) for RecSys is growing, effective defenses remain relatively underdeveloped. We categorize the existing defenses against RecSys MIAs into: proactive vs post-hoc approaches. In proactive approaches, model owners integrate privacy protection methods into RecSys modeling and system development, often at the cost of utility loss. The representative methods are differential privacy, regularization, and popularity randomization. In contrast, post-hoc mechanisms take utility as the first priority and try to meet the privacy requirements afterwards, including privacy risk estimation, and machine unlearning.

4.1 Proactive Methods

The proactive methods are dominated by differential privacy, while a few studies have also used regularization methods to address the model overfitting problem, which is believed to be the root cause of MIA.

4.1.1 Differential Privacy. Differential privacy (DP) [3] is a rigorous probabilistic mechanism that provides information-theoretic guarantees of privacy: when a machine learning model is trained under a suitably small privacy budget, it cannot reliably learn or remember any single user's data if the privacy budget is sufficiently small. The definition is that A (possibly randomized) mechanism M is said to satisfy (ε, δ)-differential privacy (DP)[3, 35, 56] if, for all pairs of adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in exactly one record, and for all measurable output events S, the following holds:

$$\Pr[M(\mathcal{D}) \in S] \;\leq\; e^{\varepsilon} \Pr[M(\mathcal{D}') \in S] + \delta. \tag{1}$$

This inequality ensures that the inclusion or exclusion of any single record has only a limited effect on the mechanism's output distribution, thereby providing a rigorous privacy guarantee. Furthermore, the MIA effectiveness level can be theoretically linked to and bounded by the privacy budget of DP, ε, as shown in later discussion, which is the unique strength of DP.

In the context of RecSys, DP has been applied to mitigate membership inference attacks (MIAs) by adding calibrated noise to inputs, gradients, or model outputs [89, 91]. Local differential privacy (LDP) is a client-side variant in which each user independently perturbs their data or model updates before transmitting them, thereby protecting against server-side inference. In federated RecSys, for instance, LDP has been evaluated as a defense strategy against interaction-level MIAs [82], showing that unless extremely large noise is applied the attacker's accuracy remains high, while excessive noise severely degrades recommendation performance. These findings highlight a fundamental trade-off: although DP and LDP offer formal membership-privacy guarantees, their practical deployment in RecSys is constrained by the privacy–utility dilemma. For example, while using LDP on a federated recommendation system reduces the effectiveness of attacks by up to 67%, it also decreases the recommendation accuracy by 65% [82]; on a centralized recommendation system, when the defense effectiveness reaches 40%, the model performance drops by 90%. Designing defense mechanisms that preserve recommendation quality while providing tangible membership-privacy protection thus remains a critical open problem.

4.1.2 Regularization. Regularization aims to reduce the overfitting degree of target models to mitigate MIAs. Therefore, regularization methods that can reduce the overfitting of ML models can be leveraged to defend against MIAs. Existing regularization methods, including L2-norm regularization, dropout [67], data argumentation, model stacking, early stopping, label smoothing [69], adversarial regularization [57], and Mixup + MMD (Maximum Mean Discrepancy) [39], have been proposed and investigated as defense methods against MIAs in other fields [7, 51, 64, 66].

However, the study of using regularization as a defense method in RecSys is limited. Zhong et al. [91] firstly proposes regularization via gradient-level learning (RGL). The key idea is to introduce a regularization term into the target RecSys training objective to reduce the distinguishability between member and non-member samples from the perspective of the surrogate attacker. The regularization term is defined using the Kullback–Leibler (KL) divergence between the probability distributions of attack predictions for members and non-members. By penalizing the discrepancy between these distributions, RGL effectively mitigates the information gap exploited by MIAs and increase the defense efficiency. For example, RGL can increase the defense efficiency up to 37.8% under MINER attack.

In the domain of social RecSys[89], memorization of training data can inadvertently expose sensitive social relationships between users via recommendation outputs. Unlike many other application areas, social RecSys relies critically on user-to-user links to enhance recommendation quality. To mitigate this risk, the concept of Socially Adversarial Learning (SAL) has been developed specifically for the recommendation field. Under SAL, a surrogate attacker $\mathcal{A}'$ is embedded into model training: the attacker learns to distinguish user pairs (u_1, u_2) that share a social link ($\mathcal{U}^+$) from those that do not ($\mathcal{U}^-$). The RecSys is then optimized with the combined objective

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{rec}} + \lambda\, \mathcal{L}_{\text{def}},$$

where

$$\mathcal{L}_{\text{def}} = \text{Dis}\big(\mathcal{A}'(\mathcal{U}^+),\, \mathcal{A}'(\mathcal{U}^-)\big) \approx \sqrt{(\mu^+ - \mu^-)^2 + (\sigma^+ - \sigma^-)^2},$$

in which μ^+, μ^- and σ^+, σ^- denote the means and standard deviations of the surrogate attacker's output distributions for linked and non-linked user pairs, respectively. By minimizing this divergence term, SAL forces user-pair embeddings and recommendation outcomes for socially linked and unlinked users to become indistinguishable — thereby reducing the risk of social-relationship leakage.

Except for centralized RecSys, regularization has been used in Federated RecSys to protect user privacy information. To defend against interaction-level MIAs in Federated RecSys, Yuan et al. [82] proposed a regularization-style update control mechanism. Observing that user embeddings change little during training and that the public parameter updates from each client can leak membership signals, the approach augments the client-side optimization objective with a penalty term:

$$\mathcal{L} = \mathcal{L}^{\text{rec}} + \mu \, \|V^t - V^0\|,$$

where $\mathcal{L}^{\text{rec}}$ denotes the standard recommendation loss, V^0 is the client's initial public parameter vector, V^t is the uploaded parameter update at round t, and μ is the regularization strength. By constraining clients' updates from drifting far from the initial state, the defense reduces the distinguishability of members versus non-members that an attacker can exploit, while preserving recommendation utility more effectively than straightforward local differential privacy (LDP) noise addition.

While regularization-based defenses have received more attention in recommendation systems than differential privacy, they still inevitably involve a trade-off between privacy protection and recommendation utility. For instance, RGL improves defense effectiveness by up to 37.8%, but this gain comes at the cost of a substantial utility drop of up to 12.7%. Similarly, SAL enhances defense efficiency by up to 6.0%, yet reduces model performance by as much as 9.7%. These results highlight a fundamental limitation of regularization-based defenses: improving robustness against membership inference attacks often comes at the expense of degrading recommendation quality. Designing regularization techniques that effectively address the MIA threat, while minimizing utility loss, remains a critical open challenge.

4.1.3 Popularity Randomization. To mitigate membership inference attacks in RecSys, Zhang et al. [84] propose a defense mechanism called Popularity Randomization. The core idea is that non-member users are typically recommended the most popular items, making their latent feature vectors unusually similar and therefore easily distinguished from members. To counter this weakness, the system expands the candidate pool of popular items for non-members and then randomly selects a subset for recommendation, thereby increasing randomness in non-members' outputs. Formally, when issuing recommendations to non-member users, instead of always selecting the top-k popular items, the method chooses a larger set of top-N popular items and randomly picks k items from within that set. This randomization breaks the deterministic mapping of non-members to the most popular items and reduces the similarity in feature vectors between non-members and members, thereby lowering the distinguishability exploited by the attack. However, this defense is vulnerable to attacks that assume non-members receive recommendations dominated by globally popular items. Modern RecSys can effectively address cold-start issues, meaning that even non-members often obtain reasonably personalized recommendations. As a result, the underlying assumption of this defense may not always hold in practice. Moreover, recent MIA studies have demonstrated that personalized recommendations themselves can leak sensitive user information [92]. Therefore, the practical feasibility of this defense strategy remains uncertain.

4.2 Post-hoc Methods

Post-hoc methods aim to preserve utility first and meet the privacy requirements later. It avoids interfering with the modeling process, and thus fully preserves model utility before the system is deployed. Under this approach, the model owner conforms the privacy laws, e.g., providing the required security measures and following the data minimization principle, and responds to users' "right to be forgotten" requests after the system is deployed. When users request removing their data from modeling, a procedure called *machine unlearning* is often used. In addition to that, the concept of *privacy risk estimation* is introduced to allow both users and the model builder to learn the privacy risk of each contributed item, which may be used to decide data items to be removed. Due to the fully preserved utility, post-hoc methods might be more accepted in practice.

4.2.1 Machine Unlearning. Unlearning has become a widely adopted post-hoc defense method, enabling model owners to meet users' privacy requests after deployment. Following the notion of unlearning principles [59], we further categorize recommendation unlearning techniques into exact unlearning and approximate unlearning, depending on whether the method fully or partially removes the influence of a user's data from the trained model.

Exact Unlearning. Exact unlearning follows a strict and complete definition of machine unlearning, aiming to fully eliminate the influence of designated data samples at the algorithmic level. Inspired by the SISA method [6], most exact recommendation unlearning methods adopt the ensemble retraining framework. This framework partitions the original dataset into multiple subsets, trains a sub-model on each subset, and aggregates these sub-models to form the final predictor—similar to an ensemble learning pipeline. To guarantee algorithmic completeness, each sub-model is typically designed to be identical to the original model in terms of architecture, hyper-parameters, and training configurations. This design enables efficient unlearning: when a user submits an unlearning request, only the sub-model trained on the subset containing the target data needs to be retrained, avoiding full retraining of the entire dataset and thereby substantially improving efficiency.

Building upon SISA, Chen et al.[9] propose RecEraser, which introduces two key modifications tailored to recommendation tasks. First, RecEraser employs a balanced clustering module for dataset partitioning, grouping similar users or items into the same subset so as to preserve collaborative effects—unlike the random partitioning strategy used in SISA. Second, RecEraser incorporates an attention-based aggregation network that learns adaptive weights for combining sub-models. Compared to uniform averaging or majority voting in SISA, this weighted aggregation significantly improves recommendation performance. Despite its advantages, the ensemble retraining framework faces a fundamental trade-off between unlearning efficiency and model utility in recommendation settings. Increasing the number of data partitions improves unlearning efficiency, but it also weakens collaborative signals, thereby degrading recommendation quality.

To preserve more utility, Li et al. [41] propose UltraRE, a lightweight extension of RecEraser. UltraRE introduces a novel balanced clustering algorithm based on optimal transport theory, improving both clustering quality and computational efficiency. Furthermore, UltraRE simplifies the attention module by replacing it with a logistic regression model, further enhancing overall efficiency.

Also, aiming to preserve more model utility, LASER adopts sequential training instead of parallel training during model aggregation [43]. Sequential training processes sub-models one after another, which helps preserve cross-subset collaborative patterns that may be lost during parallel training. LASER additionally integrates curriculum learning to optimize the ordering of data subsets, further enhancing predictive performance. However, sequential training inevitably

reduces unlearning efficiency. To address this, LASER incorporates early stopping and parameter manipulation strategies to shorten retraining time while maintaining unlearning completeness.

Exact unlearning provides strong privacy guarantees by ensuring that all traces of targeted user data are completely removed from the training process, thereby eliminating their influence on the final model.

Approximate Unlearning. Exact unlearning requires the presence of the entire training data, which is expensive to maintain and may not be available in a running system. Approximate unlearning aim to *approximate* the effect of exact unlearning. These approaches operate either from a parametric perspective, directly manipulating model parameters, or from a functional perspective, fine-tuning the model so that its behavior resembles that of a model retrained without the forgotten data.

A major class of parametric approaches is reverse unlearning[42, 85, 88], where the influence of the target data is estimated and subtracted from the model parameters. Influence-function-based methods estimate this effect using gradient and Hessian information, providing a closed-form update that avoids additional training. However, in recommendation systems with high-dimensional user–item embeddings, these methods face significant computational overhead and may suffer from estimation inaccuracies. To mitigate this, recent work selectively computes influence only for target embeddings or prunes less important parameters to reduce the cost. From a functional perspective, active unlearning methods fine-tune the model toward an unlearned solution. Representative approaches include fine-tuning on retained data[48] and using flipped losses for forget items[4]. These methods are faster than exact retraining and can be applied to an already trained model, but lack theoretical guarantees because their performance depends heavily on optimization stability and the design of the fine-tuning objective.

Overall, approximate unlearning offers greater efficiency and is more practical for large-scale RecSys, but its reliability regarding privacy protection has been questioned [19].

4.2.2 Privacy Risk Estimation. Unlearning provides a principled mechanism for ensuring user privacy by enabling the complete removal of a user's data from the training corpus. From the user's perspective, this capability aligns with the "right to be forgotten," allowing individuals to request that their personal information be permanently deleted from the model's training set. From the company's perspective, however, a central challenge lies in systematically quantifying the privacy risk associated with retaining or removing specific user data. In this context, He et al.[22] introduced the **privacy risk score** that offers a formal measure of how much sensitive information a model may inadvertently reveal about a user's data, thereby providing a quantitative foundation for evaluating and enforcing privacy guarantees in machine learning systems. The privacy risk score is inspired by the differential privacy. In Eq. (1), under the hypothesis-testing interpretation of differential privacy, let S denote the rejection region used by an arbitrary statistical test (or distinguisher) attempting to decide whether the mechanism's output originated from $\mathcal{D}$ or $\mathcal{D}'$. If TPR and FPR represent the true and false positive rates of this test, respectively, then the privacy guarantee in Eq. (1) imposes the following constraint on the achievable tradeoff between them:

$$\text{TPR} \;\leq\; e^{\varepsilon}\,\text{FPR} + \delta, \tag{2}$$

and symmetrically with the roles of $\mathcal{D}$ and $\mathcal{D}'$ swapped.

Equation (2) provides an intuitive ROC-curve interpretation of differential privacy: a smaller (ε, δ) pair uniformly bounds the distinguishing power of any potential adversary, thereby limiting the success probability of all membership inference attacks (MIAs) against the mechanism. Note that δ is often very small, and thus, can be safely removed from Eq. 2.

Interpreting Eq. (2) from a pointwise perspective naturally leads to the notion of an *empirical indistinguishability level* for a specific record z, without the DP mechnism is applied. For a fixed trained model instance, a calibrated and sufficiently powerful membership inference attack (MIA) yields empirical values of (TPR, FPR) at a chosen operating point. Rearranging Eq. (2) then motivates the definition of a sample-specific privacy score [22]:

$$\hat{\varepsilon}_z \approx \log\left(\frac{\mathrm{TPR}}{\max\{\mathrm{FPR}, \epsilon_{\mathrm{num}}\}}\right),$$

where a small ϵ_{num} ensures numerical stability. Intuitively, $\hat{\varepsilon}_z$ quantifies the empirical privacy leakage of an individual record: higher values indicate that the record is easier to distinguish between membership and non-membership, and thus more privacy-sensitive for that model instance.

In the context of RecSys, MIAs instantiate this principle at various granularities—*user-level* and *interaction-level*—by leveraging observable signals such as predicted scores, item ranks, or conversational traces. Consequently, such calibrated MIAs provide a practical, per-sample proxy for assessing privacy risk in RecSys, aligning empirical vulnerability measurement with the formal hypothesis-testing interpretation of differential privacy. The model owner can use this privacy risk score as a guiding instruction to remove the sensitive data from their training data to achieve the defensive purpose. Similar to the taxonomies of attacks, we also give readers a general picture of membership inference defenses to help readers find the most relevant papers easily. The taxonomy of membership inference defenses in illustrated in 5. In this taxonomy, we categorize all released papers of membership inference defenses into two main categories, i.e., proactive and post-hoc based defenses. For the papers under each of the categories, we further divide the papers based on the specific defense approach, enabling the readers to find the most relevant papers.

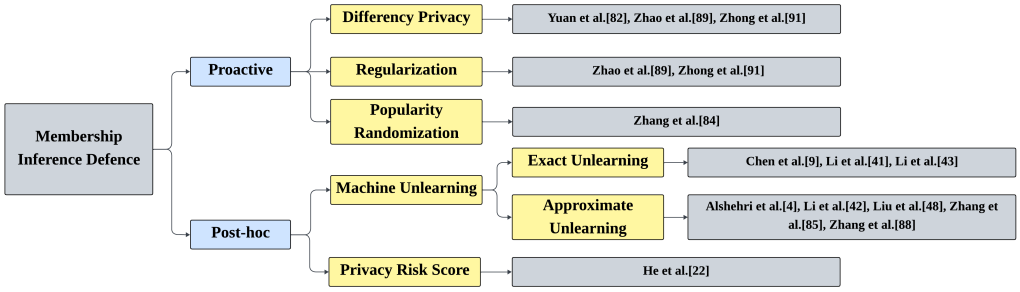

Fig. 5. Taxonomy: Membership Inference Defense on Recommendation System.

5 Evaluation

Building upon the aforementioned design principles, the evaluation of Membership Inference Attack (MIA) performance in RecSys primarily centers on three fundamental metrics: the Area Under the Receiver Operating Characteristic Curve (AUC), the F1-score, and the True Positive Rate (TPR) to False Positive Rate (FPR) ratio. Each of these metrics serves a distinct yet complementary role in characterizing the attack's discriminative capability, robustness, and real-world applicability. In this section, we present a comprehensive overview of the evaluation methodology, encompassing the benchmark datasets, representative recommendation models, and standardized performance indicators commonly adopted in empirical studies of MIAs on recommendation systems.

5.0.1 Datasets. MIA methods on RecSys use the same datasets as other recommendation tasks. We list the widely used datasets and summarize the statistics in Table 1.

MovieLens[20] The MovieLens dataset is one of the most widely adopted and benchmarked datasets in recommender system research. It comprises user–movie rating interactions and is available in multiple versions that differ in scale. The numeric suffix in each version denotes the approximate number of rating records it contains; for instance, MovieLens-1M includes about 1M user–item interactions.

Amazon[29] The Amazon dataset comprises multiple domain-specific subsets categorized according to product types available on the Amazon platform. Each sub-dataset contains user reviews and metadata related to a particular product category. For instance, ADM, Beauty, Book, and Cell Phone represent the sub-datasets corresponding to Digital Music, Beauty products, books, and communication equipment, respectively.

LastFM The LastFM dataset is a music listening dataset collected from the Last.fm online music platform. It contains user–artist interaction records, such as listening histories and play counts, and is widely used for evaluating music recommendation and user preference modeling tasks.

Steam[71] The Steam dataset was collected from Steam, one of the world's largest digital distribution platforms for PC games. It contains detailed user–game interaction records, including transaction data such as game purchases and playtime durations. Among its various releases, Steam-200K is a widely adopted subset that serves as a benchmark for evaluating recommendation models in gaming-related domains.

Ta-feng [14] The Ta-Feng dataset is a supermarket transaction dataset collected from a retail chain in Taiwan. It records detailed purchase histories, including user IDs, product IDs, quantities, and timestamps. Due to its sequential and temporal characteristics, it is widely used for sequential recommendation and next-item prediction research.

Yelp18[34] The Yelp dataset was originally compiled for the Yelp Dataset Challenge and contains users' reviews of restaurants. The company Yelp3 is a platform that publishes crowd-sourced reviews of restaurants. which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries. A cleaned subset of Yelp reviews with user–business interactions and ratings, used for recommendation and review analysis.

Ciao[72] The Ciao dataset was collected from the Ciao online product review platform, which allows users to rate, review, and socially interact with other users. It contains rich user–item interactions along with explicit social relations (e.g., trust networks), making it well-suited for evaluating social recommendation models. Each record includes user ratings, product metadata, and user trust links, enabling comprehensive analysis of both behavioral and social influence factors in recommendation tasks.

Flickr[2] The Flickr dataset was extracted from the Flickr social image-sharing platform and includes users' interactions with images (e.g., favorites, comments, or tags), as well as the underlying user–user social connections. It provides a representative benchmark for studying socially-aware recommendation systems, as it integrates both content-based and social network information, reflecting real-world scenarios where user preferences are shaped by social relationships and shared media interests.

5.0.2 Models. To verify the effectiveness of the proposed methods, MIAs are often evaluated on various popular RecSys models. We list the widely used RecSys model structures as follows:

ICF[62] calculates the similarity between items aiming to find the ones which are closed to users' likes.

LFM[37] builds a latent space to bridge user preferences and item attributes.

NCF[26] A key collaborative filtering model that leverages neural network architectures.

[2]https://www.flickr.com/

Table 1. Statistics of widely used datasets for MIAs on RecSys.

Dataset	User #	Item #	Interaction #
MovieLens-100K[20]	943	1,682	100,000
MovieLens-1M[20]	6,040	3,706	1,000,209
Yelp[34]	1,987,897	150,346	6,990,280
Ta-Feng[14]	32,266	23,812	817,742
Amazon-Digital Music[29]	100,952	70,519	130,434
Amazon-Beauty[29]	631,986	115,709	701,528
Amazon-Book[29]	10,297,355	4,493,336	29,475,453
Amazon-Cell Phone[29]	11,598,197	1,623,399	20,812,945
LastFM	23,566	48,123	1,474,122
Steam[71]	12,393	5,155	200,000
Ciao[72]	7,375	99,746	278,483
Flickr[2]	3,074,947	41,278,715	187,168,754

LightGCN[25] A state-of-the-art collaborative filtering model that simplifies graph convolution networks to enhance recommendation performance.

GRU4Rec[28] is a session-based recommendation model that uses Gated Recurrent Units (GRUs) to model a user's short-term click sequence and predict the next item. It's typically trained with pairwise ranking losses (e.g., BPR/TOP1) and negative sampling, handling variable-length sessions efficiently.

BERT4Rec[68] utilizes deep two-way transformer to model the sequence of user behavior, exhibiting excellent performance in multiple tasks.

STAMP[47] not only captures the user's general interests but also preserves the user's current preference through a new short-term memory priority.

NARM[40] consists of a global encoder and a local encoder. The latter encoder combined attention mechanism to attend large or small weights for different items.

CKE[83] Collaborative Knowledge base Embedding integrates a knowledge graph into matrix-factorization by jointly learning user/item embeddings with KG structural/semantic regularizers (e.g., TransE/semantic embeddings), improving cold-start and representation quality.

KGAT[78] Knowledge Graph Attention Network propagates user–item preferences over a KG with high-order neighbor aggregation via attention, then performs end-to-end recommendation with the KG-enhanced embeddings.

ECFAT Commonly cited as Explainable Collaborative Filtering with Attentive Transfer, it transfers item attribute/auxiliary information into CF via attention mechanisms to provide interpretable recommendations and better generalization

DiffNet++[81] A social recommendation model that iteratively diffuses user and item signals over the user–item bipartite and social graphs, capturing high-order social influence and preference propagation for improved link prediction.

DESIGN[74] A social recommender that performs denoising/self-supervised learning on social graphs (and possibly user–item graphs) to reduce social noise and enhance representations before prediction.

GDMSR[60] Graph Denoising based Multi-relational Social Recommendation leverages denoising objectives on social and interaction graphs to mitigate noisy/irrelevant relations while learning multi-relational user/item representations for recommendation.

5.0.3 Metrics. To verify the effectiveness of these MIAs, they are seen as binary classification to be evaluate. We summarize the common evaluation metric.

Attack Success Rate (ASR) ASR is defined as:

$$\text{ASR} = \frac{\#\,\text{Successful Attacks}}{\#\,\text{All Attacks}}.$$

A higher ASR indicates that the attack can more effectively distinguish members from non-members, reflecting stronger attack capability.

AUC (ROC-AUC).[79, 84, 89, 92] Area under the ROC curve; a threshold-independent summary of performance across all decision thresholds and interpretable as the probability that a randomly chosen positive is scored higher than a randomly chosen negative.

F1-score.[82] Harmonic mean of precision and recall:

$$\text{F1} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

balancing both types of errors into a single score in $[0, 1]$.

TPR / FPR.[23, 91]

$$\text{True Positive Rate(TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

TPR measures coverage of actual positives; FPR measures the rate of false alarms among actual negatives.

Advantage.[24] Advantage is a simple transformation of the attack classifier's accuracy:

$$\text{Advantage} \;=\; 2\,\big(1 - \text{Accuracy}\big).$$

Here, Accuracy $\in [0, 1]$ is the fraction of correctly classified cases (member vs. non-member). Thus, Advantage $= 2(1 - \text{Accuracy})$ equals twice the error rate: larger values indicate lower attack accuracy (i.e., stronger privacy), while smaller values indicate a more effective attack.

6 Discussion and Future Directions

In this section, we discuss several main challenges and potential research opportunities in MIAs on RecSys to inspire interested readers to explore this field.

6.1 Adversary modeling

In practical scenarios, black-box RecSys models most closely reflect the adversary's perspective, where attackers can exploit only limited external information to compromise user privacy. The existence of such feasible black-box attacks significantly undermines user trust in service providers and poses long-term risks to corporate reputation and sustainability. Therefore, while the community continues to explore potential attack strategies and quantify the privacy risks associated with black-box models, equal emphasis should be placed on developing effective defense mechanisms. Given that even the most rigorous differential privacy (DP) approaches often lead to substantial utility degradation—rendering them less attractive for real-world deployment—future research should focus on designing adaptive and lightweight defense frameworks that balance privacy protection with model performance and business applicability.

In contrast, white-box settings assume that attackers have access to a model's internal parameters, gradients, or architecture, which is rarely achievable in practice. Nonetheless, white-box analyses provide valuable insights from the

enterprise's perspective by quantifying internal privacy leakage and informing proactive protection strategies. The recently proposed notion of *privacy risk scores*, which links membership inference attacks (MIAs) with differential privacy theory, offers a promising avenue for enterprises to identify and "preemptively unlearn" privacy-sensitive users or interactions before deployment. Developing more efficient and interpretable methods for estimating such risk scores—while ensuring minimal performance compromise—represents a compelling future direction toward achieving privacy-preserving yet high-utility RecSys.

6.2 Importance of interaction-level MIAs

Unlike user-targeted MIAs, which concerns whether a particular user is included in a dataset, interaction-level privacy concerns whether specific user-item interactions (e.g., a user's click, purchase, or rating) were used in training a RecSys. This more granular form of privacy highlights a previously under-explored vulnerability in modern RecSys (e.g., RecSys on social media and e-commerce) and underscores heightened user concern when private or sensitive interactions are exposed. For instance, while merely knowing that a user is present in a system's training data may often be trivial (e.g., "someone uses Amazon"), inferring that a particular purchase or specific click-stream event was included is far more difficult and yet potentially far more revealing.

Current research on interaction-level privacy remains nascent, focusing primarily on federated RecSys, knowledge-graph-based systems, and in-context learning (ICL) recommender frameworks. Beyond conducting membership-inference attacks (MIAs) to expose private interactions, a promising future direction is the development of privacy-risk scoring mechanisms that quantify the sensitivity of individual interactions. With such scores, companies could proactively identify which user records pose high privacy risks and trigger targeted protective measures or the unlearning of recommendations. Further, more sophisticated MIAs may be designed by monitoring training signals (e.g., loss trajectories, prediction logits, confidence scores, and embedding drift) at the individual-interaction level and mapping them to privacy-risk indicators in large-language-model-powered recommendation systems.

6.3 Relationship between user-level and interaction-level MIAs

Despite substantial progress in understanding user-level privacy risks, several open questions remain. One fundamental uncertainty concerns the relationship between user-level and interaction-level privacy: is a user's overall privacy exposure simply the average of their interaction-level risks, or do certain interactions contribute disproportionately to that exposure? In particular, long-tail interactions—such as those involving niche items or infrequent behaviors—may reveal more distinctive user traits and thus contribute more to overall privacy leakage than interactions with popular items. Understanding how to quantify, weight, and aggregate these heterogeneous privacy contributions remains an open research challenge. Addressing these questions will not only advance theoretical understanding of user privacy composition but also guide the design of more precise and adaptive privacy protection mechanisms in RecSys.

6.4 New dimension of attack targets: social-level MIAs

Unlike user-level and interaction-level privacy threats, the risk posed by users' social relationships introduces a novel vantage point on the serious privacy dangers inherent in modern RecSys, especially on social media. Current work largely limits itself to detecting whether two users share a social connection (e.g., mutual following, friendship, or 'likes'). In real-world settings, however, a more meaningful dimension of social privacy lies in the private nature of those connections – for example, relationships with close friends or family members. A public celebrity follow may appear innocuous, but the inference that a user is connected to a close friend or family member potentially exposes far

greater personal risk. Thus, a promising future direction is the formulation of more granular social-relation membership inference attacks (MIAs) and the development of corresponding defenses that protect fine-grained user-pair social privacy without degrading recommendation utility.

6.5 Attacks on emerging RecSys models

While the membership inference attack (MIA) community has made considerable progress in studying privacy breaches across different RecSys models, several critical gaps remain. First, research on Federated RecSys is still in its infancy. The existing single study is far from sufficient to capture the breadth of potential vulnerabilities, motivating further exploration of privacy risks and defenses in diverse FL-RecSys architectures. In particular, understanding how user–item embeddings, communication compression, and personalized aggregation strategies influence privacy leakage remains an open question. Second, social-level privacy risks in graph-based Federated RecSys present a promising yet underexplored direction. In such settings, an adversary might exploit graph structures, message-passing mechanisms, or aggregation updates to infer social connections, collaborative behaviors, or sensitive relational patterns among users. Investigating privacy-preserving graph aggregation and robust communication protocols is therefore essential to strengthen privacy guarantees in future graph Federated RecSys.

On the centralized side, privacy research in large language model–based RecSys (LLM4Rec) and multimodal RecSys (MM-RecSys) remains limited. For LLM4Rec, examining privacy risks arising from large models' inference, generalization, and memorization capabilities is particularly important, as these models may inadvertently memorize user interactions or reveal training data through generated outputs. Meanwhile, MM-RecSys introduces heterogeneous data modalities—including images, text, and speech—which not only amplify user privacy exposure but also raise new concerns regarding intellectual property and content ownership.

Understanding these emerging privacy risks and designing principled defense mechanisms at both model and system levels are essential for building a transparent, accountable, and trustworthy RecSys. Proactively addressing such challenges will be key to fostering user trust and ensuring the sustainable development of privacy-preserving RecSys in both academia and industry.

6.6 Tradeoffs in defense methods

Designing defenses against MIAs in RecSys necessarily involves balancing privacy protection against utility (recommendation quality). Different defense strategies yield different trade-offs, and understanding these trade-offs is critical for evaluating which approaches are viable in practice.

The utility–privacy tradeoff is the fundamental constraint in privacy-preserving methods. Any method that seeks to hide which users contributed to the training data must avoid degrading the usefulness of recommendations too much – otherwise, the system loses its main purpose. Differential Privacy (DP) offers a theoretically rigorous foundation. By adding controlled noise to the model during training (or to outputs after training), DP can formally bound the privacy risk and provide a quantifiable privacy guarantee. However, this guarantee comes at a cost: the added noise typically reduces model accuracy, often substantially – especially for recommendation tasks, which are already sensitive to small perturbations. Because of this noise-induced utility loss, adopting DP in practical RecSys remains challenging. Other proactive methods, such as regularization and popularity randomization, attempt to reduce model memorization or overfitting to reduce the risk of MIA. These methods are easy to deploy, mitigate some privacy risks, and may reduce less model utility. However, they do not come with formal privacy guarantees. Their effectiveness of protection depends heavily on assumptions about attacker's prior knowledge, and may still incur non-trivial utility loss.

Post-hoc, e.g., unlearning, methods are appealing in practice. They fully preserve utility at the deployment time, while responding to users' requests later to conform privacy laws. This makes them attractive for practitioners, who cannot tolerate large accuracy drops. Recent approaches in machine unlearning argue that unlearning can mitigate MIAs while preserving more model utility than proactive methods like DP. Nevertheless, these approaches come with their own challenges: computational overhead (e.g., unlearning may require bookkeeping training data information and significant parameter adjustments; privacy risk estimation methods are still very expensive to deploy), complexity in implementation, and auditing the unlearning result [19].

7 Conclusion

Due to wide deployment of RecSys, privacy threats and leakage in RecSys can generate enormous impacts on individuals' everyday life. We present a comprehensive review of membership inference attacks on RecSys to cover the recent advances in this new research domain. We propose a taxonomy that organizes existing attacks, explain why and how they succeed under common RecSys settings, and summarize standard evaluation protocols, metrics, and defense strategies. We then discuss open challenges for both attacks and defenses and outline promising directions for future research. Our goal is to provide a coherent foundation for subsequent work on privacy in RecSys.

References

[1] 2018. California Consumer Privacy Act of 2018 (CCPA). Legislation enacted by the State of California. Available at https://oag.ca.gov/privacy/ccpa.

[2] 2021. Personal Information Protection Law of the People's Republic of China (PIPL). https://personalinformationprotectionlaw.com/. Accessed: November 4, 2025.

[3] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[4] Manal A. Alshehri and Xiangliang Zhang. 2023. Forgetting User Preference in Recommendation Systems with Label-Flipping. In *2023 IEEE International Conference on Big Data (BigData)*. 271–281. doi:10.1109/BigData59044.2023.10386603

[5] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*. ACM, 1007–1014. doi:10.1145/3604915.3608857

[6] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2020. Machine Unlearning. arXiv:1912.03817 [cs.CR] https://arxiv.org/abs/1912.03817

[7] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1897–1914.

[8] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. 2022. The Privacy Onion Effect: Memorization is Relative. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 13263–13276. https://proceedings.neurips.cc/paper_files/paper/2022/file/564b5f8289ba846ebc498417e834c253-Paper-Conference.pdf

[9] Chong Chen, Fei Sun, Min Zhang, and Bolin Ding. 2022. Recommendation unlearning. In *Proceedings of the ACM Web Conference 2022*. 2768–2777.

[10] Jiao Chen, Luyi Ma, Xiaohan Li, Nikhil Thakurdesai, Jianpeng Xu, Jason H. D. Cho, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Knowledge Graph Completion Models are Few-shot Learners: An Empirical Study of Relation Labeling in E-commerce with LLMs. arXiv:2305.09858 [cs.IR] https://arxiv.org/abs/2305.09858

[11] Xiao Chen, Wenqi Fan, Jingfan Chen, Haochen Liu, Zitao Liu, Zhaoxiang Zhang, and Qing Li. 2023. Fairly Adaptive Negative Sampling for Recommendations. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3723–3733. doi:10.1145/3543507.3583355

[12] Xiaoxiao Chi, Xuyun Zhang, Yan Wang, Lianyong Qi, Amin Beheshti, Xiaolong Xu, Kim-Kwang Raymond Choo, Shuo Wang, and Hongsheng Hu. 2024. Shadow-Free Membership Inference Attacks: Recommender Systems Are More Vulnerable Than You Thought. arXiv:2405.07018 [cs.CR] https://arxiv.org/abs/2405.07018

[13] Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems. arXiv:2205.08084 [cs.IR] https://arxiv.org/abs/2205.08084

[14] Chiranjiv Das. 2018. Ta Feng Grocery Dataset. https://www.kaggle.com/datasets/chiranjivdas09/ta-feng-grocery-dataset. Accessed: 2025-11-06.

[15] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. arXiv:2303.14524 [cs.IR] https://arxiv.org/abs/2303.14524

[16] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 299–315. doi:10.1145/3523227.3546767

[17] Yuechun Gu, Jiajie He, and Keke Chen. 2024. Demo: FT-PrivacyScore: Personalized Privacy Scoring Service for Machine Learning Participation. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA, 5075–5077.

[18] Yuechun Gu, Jiajie He, and Keke Chen. 2025. Adaptive Domain Inference Attack with Concept Hierarchy. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 413–424. doi:10.1145/3690624.3709332

[19] Yuechun Gu, Jiajie He, and Keke Chen. 2025. Auditing Approximate Machine Unlearning for Differentially Private Models. arXiv:2508.18671 [cs.LG] https://arxiv.org/abs/2508.18671

[20] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[21] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. [n. d.]. LOGAN: Membership Inference Attacks Against Generative Models. *Proceedings on Privacy Enhancing Technologies* 2019, 1 ([n. d.]), 133–152.

[22] Jiajie He, Yuechun Gu, and Keke Chen. 2025. RecPS: Privacy Risk Scoring for Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. Association for Computing Machinery, New York, NY, USA, 432–440. doi:10.1145/3705328.3748052

[23] Jiajie He, Yuechun Gu, and Keke Chen. 2025. RecPS: Privacy Risk Scoring for Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. ACM, 432–440. doi:10.1145/3705328.3748052

[24] Jiajie He, Yuechun Gu, Min-Chun Chen, and Keke Chen. 2025. Membership Inference Attacks on LLM-based Recommender Systems. arXiv:2508.18665 [cs.IR] https://arxiv.org/abs/2508.18665

[25] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[26] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[27] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based Recommendations with Recurrent Neural Networks. http://arxiv.org/abs/1511.06939 cite arxiv:1511.06939Comment: Camera ready version (17th February, 2016) Affiliation update (29th March, 2016).

[28] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *International Conference on Learning Representations (ICLR)*.

[29] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. *arXiv preprint arXiv:2403.03952* (2024).

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

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

[32] Qi Hu and Yangqiu Song. 2024. User Consented Federated Recommender System Against Personalized Attribute Inference Attack. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (Merida, Mexico) *(WSDM '24)*. Association for Computing Machinery, New York, NY, USA, 276–285. doi:10.1145/3616855.3635830

[33] Mubashir Imran, Hongzhi Yin, Tong Chen, Quoc Viet Hung Nguyen, Alexander Zhou, and Kai Zheng. 2023. ReFRS: Resource-efficient Federated Recommender System for Dynamic and Diversified User Preferences. *ACM Trans. Inf. Syst.* 41, 3, Article 65 (Feb. 2023), 30 pages. doi:10.1145/3560486

[34] Yelp Inc. 2025. Yelp Open Dataset. https://business.yelp.com/data/resources/open-dataset/. Accessed: 2025-11-06.

[35] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The composition theorem for differential privacy. In *International conference on machine learning*. PMLR, 1376–1385.

[36] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation . In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society, Los Alamitos, CA, USA, 197–206. doi:10.1109/ICDM.2018.00035

[37] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. In *Computer*, Vol. 42. IEEE, 30–37.

[38] Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. 2024. SeqMIA: Sequential-Metric Based Membership Inference Attack. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA, 3496–3510. doi:10.1145/3658644.3690335

[39] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2021. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*. 5–16.

[40] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Defu Lian, Shengxian Ma, and Maarten de Rijke. 2017. Neural Attentive Session-based Recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*. ACM, 1449–1458.

[41] Yuyuan Li, Chaochao Chen, Yizhao Zhang, Weiming Liu, Lingjuan Lyu, Xiaolin Zheng, Dan Meng, and Jun Wang. 2023. UltraRE: Enhancing RecEraser for Recommendation Unlearning via Error Decomposition. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 12611–12625. https://proceedings.neurips.cc/paper_files/paper/2023/file/29a0ea49a103a233b17c0705cdeccb66-Paper-Conference.pdf

[42] Yuyuan Li, Chaochao Chen, Xiaolin Zheng, Yizhao Zhang, Biao Gong, Jun Wang, and Linxun Chen. 2023. Selective and collaborative influence function for efficient recommendation unlearning. *Expert Systems with Applications* 234 (2023), 121025.

[43] Yuyuan Li, Xiaolin Zheng, Chaochao Chen, and Junlin Liu. 2022. Making Recommender Systems Forget: Learning and Unlearning for Erasable Recommendation. arXiv:2203.11491 [cs.IR] https://arxiv.org/abs/2203.11491

[44] Guanyu Lin, Feng Liang, Weike Pan, and Zhong Ming. 2021. FedRec: Federated Recommendation With Explicit Feedback. *IEEE Intelligent Systems* 36, 5 (2021), 21–30. doi:10.1109/MIS.2020.3017205

[45] Zhaohao Lin, Weike Pan, and Zhong Ming. 2021. FR-FMSS: Federated Recommendation via Fake Marks and Secret Sharing. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 668–673. doi:10.1145/3460231.3478855

[46] Zhaohao Lin, Weike Pan, Qiang Yang, and Zhong Ming. 2022. A Generic Federated Recommendation Framework via Fake Marks and Secret Sharing. *ACM Trans. Inf. Syst.* 41, 2, Article 40 (Dec. 2022), 37 pages. doi:10.1145/3548456

[47] Qiao Liu, Yifu Zeng, Ria Mokhosi, and Haibin Zhang. 2018. STAMP: Short-Term Attention/Memory Priority Model for Session-based Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1831–1839.

[48] Wenyan Liu, Juncheng Wan, Xiaoling Wang, Weinan Zhang, Dell Zhang, and Hang Li. 2022. Forgetting Fast in Recommender Systems. arXiv:2208.06875 [cs.IR] https://arxiv.org/abs/2208.06875

[49] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. 2022. Membership Inference Attacks by Exploiting Loss Trajectory. arXiv:2208.14933 [cs.CR] https://arxiv.org/abs/2208.14933

[50] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. 2023. SimpleX: A Simple and Strong Baseline for Collaborative Filtering. arXiv:2109.12613 [cs.IR] https://arxiv.org/abs/2109.12613

[51] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. 2023. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*. IEEE, 77–83.

[52] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html

[53] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning Differentially Private Recurrent Language Models. arXiv:1710.06963 [cs.LG] https://arxiv.org/abs/1710.06963

[54] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 691–706.

[55] Khalil Muhammad, Qinqin Wang, Diarmuid O'Reilly-Morgan, Elias Tragos, Barry Smyth, Neil Hurley, James Geraci, and Aonghus Lawlor. 2020. FedFast: Going Beyond Average for Faster Training of Federated Recommender Systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1234–1242. doi:10.1145/3394486.3403176

[56] Peter Müllner, Elisabeth Lex, Markus Schedl, and Dominik Kowald. 2023. Differential privacy in collaborative filtering recommender systems: a review. *Frontiers in big Data* 6 (2023), 1249997.

[57] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 634–646.

[58] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 739–753. doi:10.1109/sp.2019.00065

[59] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299* (2022).

[60] Yuhan Quan, Jingtao Ding, Chen Gao, Lingling Yi, Depeng Jin, and Yong Li. 2023. Robust Preference-Guided Denoising for Graph based Social Recommendation. In *Proceedings of the ACM Web Conference 2023 (WWW)*.

[61] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian Personalized Ranking from Implicit Feedback. arXiv:1205.2618 [cs.IR] https://arxiv.org/abs/1205.2618

[62] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[63] Muhammad A. Shah, Joseph Szurley, Markus Mueller, Athanasios Mouchtaris, and Jasha Droppo. 2021. Evaluating the Vulnerability of End-to-End Automatic Speech Recognition Models to Membership Inference Attacks. In *Interspeech 2021*. 891–895. doi:10.21437/Interspeech.2021-1188

[64] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[65] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[66] Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2615–2632.

[67] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[68] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[69] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[70] Elham Tabassi, Kevin Burns, Michael Hadjimichael, Andres Molina-Markham, and Julian Sexton. 2019. A taxonomy and terminology of adversarial machine learning. *Journal of Research of the National Institute of Standards and Technology* (2019), 1–29.

[71] Tamber. 2017. Steam Video Games Dataset. https://www.kaggle.com/datasets/tamber/steam-video-games. Accessed: 2025-11-06.

[72] Jiliang Tang. 2014. Trust/Distrust Computing. https://www.cse.msu.edu/~tangjili/trust.html. Accessed: 2025-11-07.

[73] Jiaxi Tang and Ke Wang. 2018. Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*. ACM Press. doi:10.1145/3159652.3159656

[74] Ye Tao, Ying Li, Su Zhang, Zhirong Hou, and Zhonghai Wu. 2022. Revisiting graph based social recommendation: A distillation enhanced social graph network. In *Proceedings of the ACM Web Conference 2022*. 2830–2838.

[75] Muhammad Ammad ud din, Elena Ivannikova, Suleiman A. Khan, Were Oyomno, Qiang Fu, Kuan Eeik Tan, and Adrian Flanagan. 2019. Federated Collaborative Filtering for Privacy-Preserving Personalized Recommendation System. arXiv:1901.09888 [cs.IR] https://arxiv.org/abs/1901.09888

[76] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.

[77] Qinyong Wang, Hongzhi Yin, Tong Chen, Junliang Yu, Alexander Zhou, and Xiangliang Zhang. 2021. Fast-adapting and Privacy-preserving Federated Recommender System. arXiv:2104.00919 [cs.IR] https://arxiv.org/abs/2104.00919

[78] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 950–958.

[79] Zihan Wang, Na Huang, Fei Sun, Pengjie Ren, Zhumin Chen, Hengliang Luo, Maarten de Rijke, and Zhaochun Ren. 2022. Debiasing learning for membership inference attacks against recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1959–1968.

[80] Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. 2024. Membership Inference Attacks Against In-Context Learning. arXiv:2409.01380 [cs.CR] https://arxiv.org/abs/2409.01380

[81] Le Wu, Junwei Li, Peijie Sun, Richang Hong, Yong Ge, and Meng Wang. 2020. DiffNet++: A Neural Influence and Interest Diffusion Network for Social Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[82] Wei Yuan, Chaoqun Yang, Quoc Viet Hung Nguyen, Lizhen Cui, Tieke He, and Hongzhi Yin. 2023. Interaction-level membership inference attack against federated recommender systems. In *Proceedings of the ACM Web Conference 2023*. 1053–1062.

[83] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 353–362.

[84] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership inference attacks against recommender systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 864–879.

[85] Shuijing Zhang, Jian Lou, Li Xiong, Xiaoyu Zhang, and Jing Liu. 2023. Closed-form Machine Unlearning for Matrix Factorization. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 3278–3287. doi:10.1145/3583780.3614811

[86] Shijie Zhang, Hongzhi Yin, Tong Chen, Zi Huang, Lizhen Cui, and Xiangliang Zhang. 2021. Graph Embedding for Recommendation against Attribute Inference Attacks. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3002–3014. doi:10.1145/3442381.3449813

[87] Shijie Zhang, Wei Yuan, and Hongzhi Yin. 2023. Comprehensive Privacy Analysis on Federated Recommender System against Attribute Inference Attacks. arXiv:2205.11857 [cs.IR] https://arxiv.org/abs/2205.11857

[88] Yang Zhang, Zhiyu Hu, Yimeng Bai, Jiancan Wu, Qifan Wang, and Fuli Feng. 2024. Recommendation Unlearning via Influence Function. *ACM Trans. Recomm. Syst.* 3, 2, Article 22 (Dec. 2024), 23 pages. doi:10.1145/3701763

[89] Xuhao Zhao, Zhongrui Zhang, Yanmin Zhu, Zhaobo Wang, Wenze Ma, Jiadi Yu, and Feilong Tang. 2025. Social Relation-Level Privacy Risks and Preservation in Social Recommender Systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 1728–1737. doi:10.1145/3726302.3730086

[90] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2024. Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Transactions on Knowledge and Data Engineering* 36, 11 (Nov. 2024), 6889–6907. doi:10.1109/tkde.2024.3392335

[91] Da Zhong, Xiuling Wang, Zhichao Xu, Jun Xu, and Wendy Hui Wang. 2024. Interaction-level Membership Inference Attack against Recommender Systems with Long-tailed Distribution. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*.

3433–3442.

[92] Zhihao Zhu, Chenwang Wu, Rui Fan, Defu Lian, and Enhong Chen. 2023. Membership inference attacks against sequential recommender systems. In *Proceedings of the ACM Web Conference 2023*. 1208–1219.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009