Analysing supervised learning approaches for detecting shilling attacks in collaborative recommendations

Mrunal Kewalram Shende [1*], Vijay Verma [1]

[1] National Institute of Technology, Kurukshetra, Haryana

Abstract. Collaborative recommendation systems offer users personalized recommendations based on their past interactions and the actions of other users. However, these systems can be compromised by shilling attacks, in which fake feedback and ratings are introduced in order to manipulate the recommendations made by the system. It is important to identify and mitigate these attacks to maintain the reliability and accuracy of the recommendations. There are different ways to deal with shilling attacks, which involve using technology to detect fake ratings and assess the trustworthiness of users. Some solutions include using machine learning to spot patterns, applying filters to weed out fake ratings, using a combination of different filtering techniques to make recommendations, and establishing reputation systems to evaluate the reliability of users and their ratings. This work provides a comprehensive overview of the current methods for detecting shilling attacks in collaborative recommendation systems, including different types of attacks and various detection approaches. It also discusses the limitations and challenges of these approaches and compares their performance.

1 INTRODUCTION

Recommender systems are a class of intelligent systems that provide personalized recommendations to users based on their preferences and past interactions with a system. These systems are widely used in various domains, such as e-commerce, social media, and entertainment, where they help users navigate large amounts of information and make decisions by suggesting relevant items or content [1].
Collaborative filtering-based systems recommend items based on the preferences of similar users, taking into account the user's past interactions and the ratings or reviews of other users with similar preferences ([2]–[4]).

* Mrunal Kewalram Shende: mrunal_32113222@nitkkr.ac.in

Shilling attacks are a growing concern in the field of recommender systems, as they can significantly impact the accuracy and effectiveness of the system. Shilling attacks involve the injection of fake or biased ratings, reviews, or interactions into the system, with the aim of manipulating its recommendations in favor of specific items, products, or services ([5]–[9])Shilling attacks can take many forms, from simple attempts to boost the ratings of a particular item to sophisticated attacks that attempt to manipulate the entire recommendation algorithm. Profile injection involves creating multiple fake user profiles with specific preferences to manipulate the recommendations. By biasing the system towards specific items or content, attackers can influence the purchasing decisions of users or promote their own products or services. Profile injection attacks can be difficult to detect; as fake user profiles can be designed to appear similar to legitimate profiles. The impact of shilling attacks can be significant, as they can lead to inaccurate or biased recommendations, which can in turn, lead to decreased user satisfaction and reduced revenue for the system. Shilling attacks can also lead to decreased trust in the system, as users may begin to doubt the accuracy and fairness of the recommendations [1]

Detecting and preventing shilling attacks is a challenging task, as attackers can use sophisticated techniques to evade detection. However, various defense mechanisms have been proposed to mitigate the impact of shilling attacks.

Shilling attack detection typically involves analyzing user behavior and interaction data to identify patterns or anomalies that may indicate the presence of shilling attacks. The goal is to distinguish between legitimate user behavior and behavior that is the result of shilling attacks. Table. 1 illustrates the different machine learning techniques, including supervised, semi-supervised, and unsupervised algorithms, that have been suggested for detecting shilling attacks.

Table 1. Type of Shilling Attack Detection Techniques.

Supervised: This is a technique where the algorithm is trained on a labeled dataset, where the target variable (in this case, the presence of a shilling attack) is already known. The algorithm learns to predict the target variable based on input features (such as user ratings, item attributes, etc.) ([6], [9]–[12]).
Unsupervised: This is a technique where the algorithm is trained on a dataset that does not have any labeled data. The algorithm learns to identify patterns and relationships in the data without any prior knowledge of the target variable.
Semi-Supervised: This is a technique where the algorithm is trained on a dataset that contains both labeled and unlabeled data. The algorithm learns to use the unlabeled data to improve its predictions on the labeled data([10]–[13]).

Effective shilling attack detection is essential for ensuring the reliability and trustworthiness of recommender systems. Without effective detection mechanisms, shilling attacks can lead to inaccurate or biased recommendations, decreased user satisfaction, and reduced revenue for the system. As such, shilling attack detection is an important area of research in the field of recommender systems, and ongoing efforts are being made to develop more accurate and robust detection techniques.

2 PRELIMINARIES

2.1 Attack model

The work in [10] defined an attack profile shown in Figure 1. The attacker uses this attack profile to perform a shilling attack to manipulate the Collaborative Filtering Recommender Systems(CFRSs).
The attack profile essentially includes a set of filler items (IF), selected items (Is), a set of targeted items (IT), set of unrated items (IΦ) in case of multi-attack. It is a rantings vector of k dimensions, in which k is the total number of items in the system.

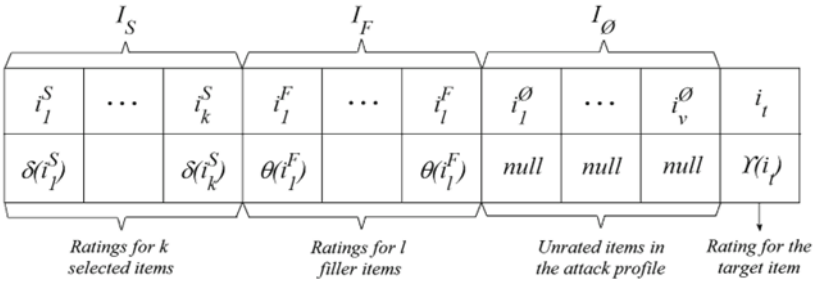

Fig.1. The General Model of Attack Profile for a Single Item[10].

2.2 Shilling Attacks Intent

- As shown in Figure 2. Push and nuke attacks are two types of attacks that can be carried out against recommender systems. These attacks are designed to manipulate the recommendations provided by the system to benefit certain items or to harm others.

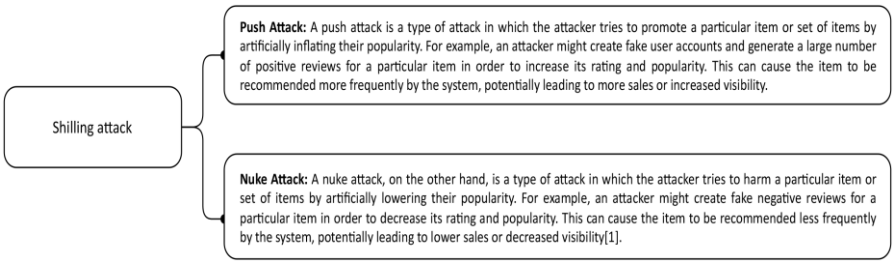

Fig.2. Two Intents of Shilling Attack.

Some prevalent attacks are explained briefly in Table 2.

Table 2: Type of Shilling Attacks.

Random Attack	In a push or nuke situation, the target item is allocated the maximum or lowest rating. [11]. The knowledge required is minimal and comes under both Nuke and Push attacks.
Average Attack	An average attack requires little more knowledge than the Random Attack, this attack is intended to perform both nuke and push attacks[12]. For each filler item average rating of all the ratings for that particular item is calculated, and that particular filler item is filled[12]

Segment Attack	These attacks[13] are often targeted at a particular item or group of items, and the fake feedback and ratings are designed to increase the perceived value or popularity of the targeted item(s) among the targeted group of users.
Bandwagon Attack	In this type of attack, An attacker creates an attack profile that contains those items that have high visibility([14][15]). The goal is to minimize the attack size by just targeting only those which occur frequently.
Love/Hate Attack	the knowledge required to perform this type of attack is shallow or equal to no knowledge required[16]. The target item receives the lowest rating, while the remaining filler objects receive the highest rating.
Probe attack	A probe attack is a type of shilling attack in which an adversary introduces fake feedback and ratings with the intent of gathering information about the system or its users.([17], [18]).

2.3 Evaluation Matric

To evaluate the performance of algorithms, several performance metrics can be used. Here are some commonly used performance metrics for shilling attack detection algorithms :
Precision : Precision is the ratio of true positives to the total number of instances classified as fake.
Accuracy : Accuracy is the ratio of correctly classified instances to the total number of instances. Accuracy measures the percentage of correctly identified fake accounts, ratings.
Recall: Recall is the ratio of true positives to the total number of actual fake instances.

3 LITERATURE REVIEW

The paper by [8] provides a comprehensive overview of shilling attacks against recommendation systems and the methods that have been proposed for detecting and preventing these attacks. The authors categorize shilling attacks into three main types: Sybil attacks, spamming attacks, and trust attacks. They also discuss various methods for detecting and preventing these attacks, including machine learning, network-based approaches, and reputation-based approaches. The paper concludes that shilling attacks can have significant negative impacts on the effectiveness and credibility of recommendation systems and that further research is needed to develop more effective methods for detecting and preventing these attacks. This work is relevant to our survey as it provides valuable insights into the nature and impact of shilling attacks and the methods that have been proposed to detect and prevent them.
While the work in [8] and [1] discuss various methods for detecting and preventing shilling attacks, they do not specifically focus on collaborative filtering recommender systems. In contrast, our survey specifically focuses on shilling attack detection in collaborative filtering recommender systems. Further, this work focuses specifically on supervised learning methods. Here, the comparative analysis provides a more targeted and specific analysis of shilling attack detection in collaborative filtering recommender systems, which can provide valuable insights for researchers and practitioners working in this area.

It has been demonstrated that collaborative filtering recommender systems are susceptible to shilling attacks([9], [11], [19], [20]). By using two ways, we can reduce the impact of a shilling attack; the first one is to perform shilling attack detection and remove the shilling

profile from the rating dataset that impacts the recommendation to the user before running the Collaborative filtering algorithm, and other way is to make the Collaborative filtering algorithm robust, i.e., an attack-resistant collaborative filtering algorithm.

A collection of labels or categories is known beforehand in supervised classification [21], and we possess a collection of labelled samples that will be utilized in a training set. The task of the classifier is to learn from the training data. Furthermore, apply the information acquired utilizing accurate data. Mapping among a feature space and a label space, where the feature provides the qualities of a classifiable entity, and the label describes the classes.

Let us now look at the contents of table.3, which lists the most important studies done on this issue.

Table 3 : Shilling Attack Detection Using Supervised Classification Techniques.

Author	Hypothesis	Targeted traits	Competent against	Weak againts	Downsides
[9]	Attack profiles are quite similar among their neighbours.	Rating Behaviour	Segment, Random	Obfuscated, Average	Misclassification of authentic user
[11]	Items with fewer ratings have higher importance	Rating Behaviour	Random, segment	Obfuscated, Average	Misclassification of authentic user
[12]	The attack type is known.	Attack size Profile Similarity	Bandwagon, Random	Obfuscated	Being attack-specific means missing out on altemative attacks.
[24]	A rating anomaly exists in attack profiles.	Filler rating variance	Bandwagon, Random, Segment	Obfuscated	Eminently susceptible on the choice of a classifier.
[25]	Selecting the pattern of a nonnal user and an attacker is different.	Item popularity and popularity distribution of user profile, rating pattern.	Average, random, Bandwagon,Segment, low-knowledge attacks.	High knowledge attack, for example, Sampling attack.	Precision value drops when a good high-knowledge attack is performed.
[26]	Weighted observations simulate a balanced dataset	Filler rating variance, filler rating length	Bandwagon attack, PIA, PUA	Probe	procedure must bc repeated several times.
[27]	There are an equal number of attack profiles authentic profiles.	Target analysis, rating behavior	Random, average, bandwagon	Obfutscated, Hybrid attack	To achieve decent outcomes, attack and authentic files must be balanced.
[28]	The detection method based on the multiple-view information gives better results.	Profile length and Filler rating variance.	Average, Random, AOP, shifting attacks.	-	Implementing and assessing every classification algorithm is tune-consuming.
[29]	Combining implicit explicit features balances the effectiveness.	User rating, Item Similarity Offset(1SO), Rating Prediction Error(RPE)	Random, Average, Bandwagon.	Detection fails when the attack is size small(l%)	It requires a huge number of training samples

[11][22]Gave the concept of generic attributes derived from each profile and used to detect the malicious user. Two trendy derived attributes came into the picture; 1. Rating Deviation from Mean Agreement(RDMA) and 2. Degree of Similarity with Top neighbors (DigSim) [9]. As an alternative, thinking of these measurements as a classification attribute, an author [11] established two new attributes employing RDMA, which are Weighted Deviation from Mean Agreement(WDMA) and Weighted Degree of Agreement(WDA). RDMA is the foundation of WDMA. [11] offer one additional general feature, length variance(lengthVar), in extension to RDMA-stationed attributes. After Further study, it was found that generic attributes alone were insufficient because of the smaller filler size usage

at that time. So Authors [11] designed attack model-specific attributes. But when the characterization of multiple profiles came, the existing techniques could not perform the classification. So, another class of attributes came into the picture. which are called intra profile detection attributes [23]. Classifiers like as SVM, kNN and C4.5 routinely employed all of the aforementioned characteristics to detect the attack. [23] Has suggested a Hybrid Attack Detection Model that is suitable for both classification approaches and statistical using Model-Specific Attributes.

Williams et al. [24] used three ways to improve detection performance in supervised systems. Similarities to reverse-engineered attacks, identification of rating anomalies, and target concentration. Author demonstrates that combining several features boosts the performance of the classier. Particularly effective is the support vector machine(SVM), which substantially decreases the effects of most aggressive attack models. Their technique employs the following attributes: WDMA, RDMA, DegSim, MeanVar, FAC, FMD, LengthVar, and FMTD. [25] proposed a supervised shilling attack detection technique. Where they proposed that the multi-level classifiers can improve efficiency. Targeted traits were profile similarity and length of the profile. His approach is effective against random, average, and bandwagon attacks. But weak against Hybrid and Obfuscated types of attacks. The approach's major downside is that the two-level classifier takes longer training time.

Another unique technique is popSAD, which was proposed in [26]. This method extracts features from the user selection pattern to detect potential shilling attacks. It is effective for any type of attack where attackers do not have full access to the database. Another technique proposed in [27] assumes that weighted observations simulate a balanced dataset, and targets filler rating variance and length to detect positive injective attacks (PIA), positive uninjective attacks (PUA), and Bandwagon attacks. However, it performs weakly against Probe attacks and requires repetitive execution. SVM-TIA, proposed in [28], is another approach that uses support vector machine classification and target item analysis to identify potential attack profiles. The method addresses class imbalance in the training data using the Borderline-SMOTE method and has a high precision rate and relatively low false positive rate. Lastly, [29] proposes the ensemble detection approach, which extracts features from ratings, user-to-user, item popularity, and sand graphs. The method employs stacked denoising autoencoders and PCA to intelligently extract user features with varied degrees of manipulation. It also uses the uniqueness of the items and the degree of variance between them as characteristics. The method follows a three-stage procedure that includes data pre-processing, feature extraction, and detection via feeble classifiers.

A unique shilling attack detection model is presented by [30] author by combining hypergraph spectral features (SpDetector). To balance performance and flexibility and handle high-order interactions by embedding hypergraphs, the proposed model mixes explicit and implicit characteristics. This technique uses rating prediction errors and item similarity offsets to identify attackers. These features are used to train a deep neural network to recognize shilling attacks.

Conclusion

In Conclusion, Shilling attacks in collaborative filtering recommender systems can have a significant impact on the reliability of recommendations. Detecting and mitigating these attacks requires a multi-faceted approach that leverages various techniques such as analysing user behaviour patterns, utilizing machine learning algorithms, and implementing reputation systems. To address the ethical implications of shilling attacks, designers and operators of collaborative filtering systems must prioritize the security and integrity of their recommendations. By effectively detecting and mitigating shilling attacks, it is possible to

prevent harm to both individuals and businesses, maintain trust in the system, and contribute to the overall success of collaborative filtering recommender systems.

References

[1] F. Rezaimehr and C. Dadkhah, "A survey of attack detection approaches in collaborative filtering recommender systems," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 2011–2066, 2021.

[2] J. Ben Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The adaptive web*, Springer, 2007, pp. 291–324.

[3] M. Nilashi *et al.*, "Collaborative filtering recommender systems," 2013.

[4] M. Elahi, F. Ricci, and N. Rubens, "A survey of active learning in collaborative filtering recommender systems," *Comput. Sci. Rev.*, vol. 20, pp. 29–50, 2016.

[5] K. Patel, A. Thakkar, C. Shah, and K. Makvana, "A state of art survey on shilling attack in collaborative filtering based recommendation system," in *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*, 2016, pp. 377–385.

[6] H. Cai and F. Zhang, "Detecting shilling attacks in recommender systems based on analysis of user rating behavior," in *Knowledge-Based Systems*, vol. 177, Elsevier, 2019, pp. 22–43.

[7] H. Yu, S. Yuan, Y. Xu, R. Ma, D. Gao, and F. Zhang, "Group attack detection in recommender systems based on triangle dense subgraph mining," in *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, IEEE, 2021, pp. 649–653.

[8] I. Gunes, C. Kaleli, A. Bilge, and H. Polat, "Shilling attacks against recommender systems: a comprehensive survey," in *Artificial Intelligence Review*, vol. 42, no. 4, Springer, 2014, pp. 767–799.

[9] P.-A. Chirita, W. Nejdl, and C. Zamfir, "Preventing shilling attacks in online recommender systems," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*, 2005, pp. 67–74.

[10] R. Bhaumik, C. Williams, B. Mobasher, and R. Burke, "Securing collaborative filtering against malicious attacks through anomaly detection," in *Proceedings of the 4th workshop on intelligent techniques for web personalization (ITWP'06), Boston*, vol. 6, 2006, p. 10.

[11] R. Burke, B. Mobasher, C. Williams, and R. Bhaumik, "Classification features for attack detection in collaborative recommender systems," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 542–547.

[12] S. K. Lam and J. Riedl, "Shilling recommender systems for fun and profit," in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 393–402.

[13] B. Mobasher, R. Burke, C. Williams, and R. Bhaumik, "Analysis and detection of segment-focused attacks against collaborative recommendation," in *International Workshop on Knowledge Discovery on the Web*, Springer, 2005, pp. 96–118.

[14] N. J. Hurley, M. P. O'Mahony, and G. C. M. Silvestre, "Attacking recommender systems: A cost-benefit analysis," in *IEEE Intelligent Systems*, vol. 22, no. 3, IEEE, 2007, pp. 64–68.

[15] R. Burke, B. Mobasher, and R. Bhaumik, "Limited knowledge shilling attacks in collaborative filtering systems," in *Proceedings of 3rd international workshop on intelligent techniques for web personalization (ITWP 2005), 19th international joint conference on artificial intelligence (IJCAI 2005)*, 2005, pp. 17–24.

[16] C. A. Williams, "Thesis: Profile injection attack detection for securing collaborative recommender systems," Citeseer, 2012.
[17] B. Mobasher, R. Burke, R. Bhaumik, and J. J. Sandvig, "Attacks and remedies in collaborative recommendation," in *IEEE Intelligent Systems*, vol. 22, no. 3, IEEE, 2007, pp. 56–63.
[18] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre, "Recommender systems: Attack types and strategies," in *AAAI*, 2005, pp. 334–339.
[19] I. Gunes, A. Bilge, C. Kaleli, and H. Polat, "Shilling attacks against privacy-preserving collaborative filtering," *J. Adv. Manag. Sci.*, vol. 1, no. 1, pp. 54–60, 2013.
[20] I. Gunes, A. Bilge, and H. Polat, "Shilling attacks against memory-based privacy-preserving recommendation algorithms," *KSII Trans. Internet Inf. Syst.*, vol. 7, no. 5, pp. 1272–1290, 2013.
[21] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging technology in modelling and graphics*, Springer, 2020, pp. 99–111.
[22] R. Burke, B. Mobasher, C. Williams, and R. Bhaumik, "Detecting profile injection attacks in collaborative recommender systems," in *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)*, 2006, p. 23.
[23] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, "Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness," in *ACM Transactions on Internet Technology (TOIT)*, vol. 7, no. 4, ACM New York, NY, USA, 2007, pp. 23-es.
[24] C. Williams and B. Mobasher, "Profile injection attack detection for securing collaborative recommender systems," *DePaul Univ. CTI Tech. Rep.*, pp. 1–47, 2006.
[25] F. Zhang and Q. Zhou, "A Meta-learning-based Approach for Detecting Profile Injection Attacks in Collaborative Recommender Systems.," *J. Comput.*, vol. 7, no. 1, pp. 226–234, 2012.
[26] W. Li, M. Gao, H. Li, J. Zeng, Q. Xiong, and S. Hirokawa, "Shilling attack detection in recommender systems via selecting patterns analysis," *IEICE Trans. Inf. Syst.*, vol. 99, no. 10, pp. 2600–2611, 2016.
[27] Z. Yang, L. Xu, Z. Cai, and Z. Xu, "Re-scale AdaBoost for attack detection in collaborative filtering recommender systems," *Knowledge-Based Syst.*, vol. 100, pp. 74–88, 2016.
[28] W. Zhou, J. Wen, Q. Xiong, M. Gao, and J. Zeng, "SVM-TIA a shilling attack detection method based on SVM and target item analysis in recommender systems," *Neurocomputing*, vol. 210, pp. 197–205, 2016.
[29] Y. Hao, F. Zhang, J. Wang, Q. Zhao, and J. Cao, "Detecting shilling attacks with automatic features from multiple views," *Secur. Commun. Networks*, vol. 2019, 2019.
[30] H. Li, M. Gao, F. Zhou, Y. Wang, Q. Fan, and L. Yang, "Fusing hypergraph spectral features for shilling attack detection," *J. Inf. Secur. Appl.*, vol. 63, p. 103051, 2021.