A Regression Framework to Interpret the Robustness of Recommender Systems Against Shilling Attacks

Discussion Paper

Yashar Deldjoo[1], Tommaso Di Noia[1], Eugenio Di Sciascio[1] and Felice Antonio Merra[1,2]

[1]Politecnico di Bari, via Orabona, 4, 70125 Bari, Italy

[2]The authors are in alphabetical order. Corresponding author: Felice Antonio Merra (felice.merra@poliba.it).

Abstract

Collaborative filtering recommender systems (CF-RSs) employ user-item feedback, e.g., ratings, purchases, or reviews, to harmonize similarities among customers and produce personalized lists of products. Being based on the benevolence of other customers, CF-RSs are vulnerable to Shilling Attacks, i.e., fake profiles injected on the platform by adversaries to hack the recommendation outcomes toward a corrupt behavior. While mainly works on shilling attacks have been conducted to propose novel methods, compare recommendation models and outputs with and without defenses, we have found a lack of study on the impact of dataset properties on the CF-RSs robustness. In this work, we present a regression model to test whether dataset characteristics can impact the robustness of CF-RSs under shilling attacks to interpret their efficacy depending on these characteristics. Obtained results can help the system designer understand the cause of CF-RSs performance variations in attack scenarios.

Keywords

Recommender systems, Shilling Attacks, Robustness

1. Introduction and Motivation

Collaborative filtering recommender systems (CF-RSs) are a core service in online platforms in increasing traffic and promoting sales [1, 2]. A key assumption of collaborative models is that users with similar preferences will likely agree to interact with novel (next) items. However, CF-RSs are vulnerable to adversarial attacks [3] such as the injection of fake profiles, named Shilling Profiles [4, 5], perturbed side-data [6, 7], or perturbed parameters [8]. The motivation for such attacks is often malicious, e.g., economic gain, market infiltration, and even for causing damage on an underlying system (break the model availability). For instance, fake social media accounts might be created to spread fake news, or false reviews could be provided about a product to promote (push) or demote (nuke) items. For instance, past works have shown that a few fake profiles (e.g., 3%) are sufficient for a prediction shift up to 30% [9, 10].

Three main directions have been explored on shilling attacks: attack designs, detection algorithms, and defense strategies. The main shilling attack strategies are random, average, popular, bandwagon, and love-hate [11]. These assume a certain level of knowledge of the adversary on the recommendation model, recommendation outputs, the properties of items

IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

yashar.deldjoo@poliba.it (Y. Deldjoo); tommaso.dinoia@poliba.it (T. Di Noia); eugenio.disciascio@poliba.it (E. Di Sciascio); felice.merra@poliba.it (F. A. Merra)

(e.g., rating mean and entropy [12])) and users (e.g., group of users [13]). Detection strategies aim to filter out fake profiles before used for the model learning [14, 15]. Robust algorithms try to reduce the influence of possible out-of-distribution profiles [16, 10].

While previous works have been orientated to "*win-lose*" scenarios, i.e., find an answer to questions such as "*Which attack models impact more the performance of specific recommendation models?*", "*Which amount of knowledge on a specific recommendation-model is required for specific attack A to influence recommendation algorithm B?*". No effort has been made to provide an interpretation on which dataset features can impact the effectiveness of attacks. Indeed, while it is well-known that CF-RSs are affected by the *sparsity of the dataset* (e.g., a denser dataset can make easier the recommendation task [17]), there are no claims in the case of shilling attacks.

In this works, we focus on a novel research question "*Given popular shilling attack types and CF models already recognized by the community, which dataset characteristics can explain an observed change in the performance of recommendation*?" To answer this question, we propose a regression-based model to analyze the effects of dataset characteristics on the robustness of CF-RSs, and, via a large-scale experiment on three domains, we evaluate how three classes of data characteristics —rating structure, rating value, and rating distribution— may influence the robustness of CF-RSs. This work is an extended abstract of [18] published at SIGIR 2020.

2. Model

Let U and I denote a set of users and items in a system, and $R \in \mathbb{R}^{|U|\times|I|}$ as the complete user-item rating matrix, where each entry $r_{ui} \in \mathbb{R}$ represents a rating assigned by user $u \in U$ to item $i \in I$ if it is a recorded interaction (we use K to indicate the set of recorded interactions), a shilling attack consists in adding novel users as composed by I_S the selected item set, I_F the filler set, I_ϕ the unrated-item set, and I_T the target item set. I_S contains items identified by the attacker to exploit the owned knowledge to maximize the effectiveness of the attack, I_F holds randomly selected items for which rating scores are assigned to make the attack imperceptible. I_ϕ includes items without ratings in the fake user profile, and I_T is the item is to push or nuke. The SP composition varies based on attack strategies. We study: **Random** [12], **Love-Hate** [19], **Bandwagon** [20], **Popular** [21], **Average** [12], and **Perfect Knowledge** [22]. To study the impact of characteristics on the efficacy of this class of attacks, we use an explanatory model defined as follows:

Definition 1 (Framework)**.** *Let D bet the set of datasets, let C be the number of data characteristic factors, let* $\mathbf{X}_c$ *be the matrix containing the independent variables values (data characteristic values specified below), then the regression model is built using the formulation*

$$\mathbf{y} = \boldsymbol{\epsilon} + \theta_0 + \boldsymbol{\theta}_c \mathbf{X}_c \tag{1}$$

where θ_0 *represents the expected value of* $\mathbf{y}$ *(the* ***attack performance metric*** *under analysis),* $\boldsymbol{\theta}_c = [\theta_1, \theta_2, ..., \theta_C]$ *is the vector of the regression coefficient associated with the IVs, and* ϵ *the error.*

Independent Variables (IV) We explore three class of independent variables on the (i) structure (i.e., $SpaceSize_{log}$,$Shape_{log}$, and $Density_{log}$), (ii) rating frequency (i.e., $Gini_{item}$ and $Gini_{user}$), and (iii) rating values (Std_{rating}) of the user-item rating matrix. F $SpaceSize_{log}$ big values might

imply a higher chance of finding similar neighbor users or items. Therefore, as both attack and recommendation models rely on identifying like-minded users (neighbor users) or similarly rated items (neighbor items), we deem $SpaceSize_{log}$ to be an impactful characteristic on evaluating the performance of shilling attacks. $Shape_{log}$ can impact the effectiveness of shilling profile injection attacks. For example, in domains where $|U| \ll |I|$) there are more candidate neighbor users than candidate neighbor items for memory-based CF models. This situation might work to the advantage of user-based CF than item-based CF. Moreover, under a similar number of ratings, changing the shape implies changing the average number of ratings per item $|K| \div |I|$. We conjecture that this circumstance may impact the robustness of CF, since the construction of SP is mainly based on altering the popularity of targeted items. $Density_{log}$ is a well-recognized issue in the community of RS and $Density_{log} = 1 - sparsity$. Sparser data means that the fraction of unrated items significantly exceeds the fraction of rated ones [23]. It can harm the performance of CF, reducing, for instance, the chance of discovering neighbors in memory-based CF, building accurate model-based CF [24]. In [25], we have already identified a potential impact of dataset density on the effectiveness of shilling attacks. $Gini_{item}$ and $Gini_{user}$ measure the concentration of items, or users', ratings and use them to capture the rating frequency distribution. The equal popularity (e.g., all users give the same number of ratings) is represented with the value of the Gini coefficients to 0, while the total inequality (e.g., only one user has given all ratings) is represented with the value to 1. Finally, we study Std_{rating} motivated by the connection between high rating variance and recommendation performance claimed in Herlocker et al. [26] and the linear and negative impact on the accuracy shown in [17].

Dependent Variables The dependent variable (DV) used to study the effectiveness of the attack on RS is the Incremental Overall Hit Ratio ($\Delta_{HR@k}$). This is a stability metric that measures if the recommendation model recommends different products due to the attack irrespective of their actual rating value [22]. The HR metric has been proposed by Aggarwal [27]

3. Experiments

We study three datasets: **ML-20M** [28] having movies ratings ($U = 138,493$, $I = 26,744$, $K = 20,000,263$, $density = 0.0054$), **Yelp** [29] containing users' reviews on businesses ($U = 25,677$, $I = 25,778$, $K = 705,994$, $density = 0.0010$), and **LFM-1b** [30] presenting user-artist play counts($U = 120,175$, $I = 521,232$, $K = 25,285,767$, $density = 0.0004$). We use three CF-RSs available in [31]: **User-*k*NN** [32], **Item-*k*NN** [32], and **SVD** [33]. Additional reproducibility details are available in the original work [18]. Table 1 presents a snapshot of the full results for answering two research questions presented below.

RQ1. Is there an underlying relationship between the studied IVs and the DV? The results obtained for the adjusted coefficient of determination ($adj.R^2$) show that the six dataset characteristics can explain more than 60% of the variation in $\Delta_{HR@k}$ irrespective of the attack type, model, and dataset, providing empirical evidence supporting the hypothesis that the six IVs can explain a large part of the DV. The explanatory power is highest for the model-based SVD approach (when comparing the global behavior of each CF model). However, not a similar observation could be made on attacks.

RQ2. How do IVs impact the DV in terms of the significance and directionality? The

Table 1
Regression results for the within dataset analysis (attack size 1%). Full table results in [18].

$\Delta_{HR@10}$		User-*k*NN			Item-*k*NN			SVD		
		ML-20M	**Yelp**	**LFM-1b**	**ML-20M**	**Yelp**	**LFM-1b**	**ML-20M**	**Yelp**	**LFM-1b**
Random	$R^2(adj.R^2)$	0.761(0.758)	0.838(0.835)	0.673(0.668)	0.820(0.818)	0.815(0.812)	0.666(0.662)	0.843(0.841)	0.908(0.907)	0.790(0.788)
	Constant	.179***	.609***	.717***	.262***	.610***	.715***	.482***	.524***	.688***
	$SpaceSize_{log}$	-0.063***	.041	-0.629***	.008	.003	-0.520***	.040*	.368***	-0.368***
	$Shape_{log}$	.184***	.248***	.288*	.139***	.198***	.125	.207***	.275***	.192
	$Density_{log}$	-0.189***	-0.316*	-1.546***	-0.174***	-0.376**	-1.366***	-0.274***	.393***	-1.047***
	$Gini_{users}$	.277	-0.012	1.901***	-0.223	.030	.891	.178	-0.660**	.988*
	$Gini_{item}$	-0.102	-0.485	1.753***	-0.305	-0.241	1.784***	.102	-1.270***	1.355***
	Std_{rating}	-0.072	.287	-0.152	-0.120	.326	.012	-0.240	.311*	-0.108
Average	$R^2(adj.R^2)$	0.759(0.756)	0.831(0.829)	0.673(0.668)	0.819(0.816)	0.813(0.811)	0.666(0.661)	0.845(0.843)	0.910(0.909)	0.790(0.788)
	Constant	.187***	.609***	.717***	.276***	.608***	.715***	.502***	.523***	.690***
	$SpaceSize_{log}$	-0.063***	.048	-0.632***	.018	.010	-0.513***	.046**	.373***	-0.339***
	$Shape_{log}$	.182***	.260***	.291*	.136***	.201***	.114	.189***	.273***	.167
	$Density_{log}$	-0.189***	-0.290*	-1.553***	-0.162***	-0.359**	-1.352***	-0.271***	.405***	-0.991***
	$Gini_{users}$	.296	.074	1.907***	-0.265	.028	.857	.095	-0.652**	.833*
	$Gini_{item}$	-0.072	-0.522	1.755***	-0.284	-0.243	1.796***	.258	-1.267***	1.317***
	Std_{rating}	-0.065	.299	-0.150	-0.114	.312	.019	-0.242	.322*	-0.079

$^{***}p \leq .001$, $^{**}p \leq .01$, $^{*}p \leq .05$

significance of the computed regression coefficients for the IVs tends to vary for each IV or group of IVs. The results show that the regression coefficients computed for $SpaceSize_{log}$, $Shape_{log}$, and $Density_{log}$ are statistically significant. This shows enough statistical evidence to support the hypothesis that structural characteristics can explain DV variations($p < 0.05, 0.01, 0.001$). However, results for the other IVs vary depending on <attack, CF-model, dataset> triplet, or they can be insignificant as in the case of Std_{rating}. For instance, the coefficients for Gini indices (i.e., $Gini_{user}$ and $Gini_{item}$) are most significant for shilling attacks against SVD, particularly for samples drawn from the Yelp and LFM-1b datasets. The coefficients for Std_{rating} are insignificant ($p\text{-}value > 0.05$) in all experimented cases, except for two cases <Random/Average attack, SVD, Yelp>, implying that this dataset characteristic, which deals directly with rating values, plays a less significant role on the impact of the attack. Investigating the *directionality* of the coefficients, Table 1 shows that the $Density_{log}$ has a **negative** effect on $\Delta_{HR@k}$ across majority of the cases in <attacks, CF-model, dataset>. This result is consistent with RSs findings that increasing the density is suitable for the performance of CF-RSs [34, 17]. An explanation is that: if we fix the number of users and items and increase the number of genuine ratings, the accuracy of similarities is improved by using more genuine ratings. As these similarities are generally vulnerable to the insertion of fake profiles, adding more genuine feedbacks can help to decrease the impact of attacks. Additionally, we can note that $SpaceSize_{log}$ has a negative impact on $\Delta_{HR@k}$ in neighborhood models, which means that increasing the space size makes neighborhood models less vulnerable to attacks. Finally, and on the contrary, $Shape_{log}$ presents a consistently positive influence on the efficacy of the attacks. We explain it by considering the following example: increasing $Shape_{log}$ leads to an increased number of users with respect to items (i.e., decreasing items). In this way, it could be easier to push the target item to higher positions inside the recommendation list (i.e., fewer items have contributed).

4. Conclusion and Future Work

We have provided statistical evidence to accept the hypothesis that the chosen properties account for a considerable portion of variations in attack performance. In particular, structural properties (i.e., size, shape, and density) have a significant impact on the model, distributional (i.e., Gini index) have a higher impact on memory-based models, and standard deviation does

not show an effect. Novel characteristics, attacks, and models are possible extensions.

Acknowledgments

We acknowledge support of PON ARS01_00876 BIO-D, Casa delle Tecnologie Emergenti della Città di Matera, PON ARS01_00821 FLET4.0, PIA Servizi Locali 2.0, H2020 Passapartout - Grant n. 101016956, and PIA ERP4.0.

References

[1] C. A. Gomez-Uribe, N. Hunt, The netflix recommender system: Algorithms, business value, and innovation, ACM Trans. Management Inf. Syst. 6 (2016) 13:1–13:19. URL: https://doi.org/10.1145/2843948. doi:10.1145/2843948.

[2] B. Smith, G. Linden, Two decades of recommender systems at amazon.com, IEEE Internet Computing 21 (2017) 12–18. URL: https://doi.org/10.1109/MIC.2017.72. doi:10.1109/MIC.2017.72.

[3] Y. Deldjoo, T. D. Noia, F. A. Merra, A survey on adversarial recommender systems: From attack/defense strategies to generative adversarial networks, ACM Comput. Surv. 54 (2021) 35:1–35:38.

[4] I. Gunes, C. Kaleli, A. Bilge, H. Polat, Shilling attacks against recommender systems: a comprehensive survey, Artif. Intell. Rev. 42 (2014) 767–799.

[5] V. W. Anelli, Y. Deldjoo, T. Di Noia, E. Di Sciascio, F. A. Merra, Sasha: Semantic-aware shilling attacks on recommender systems exploiting knowledge graphs, in: 17th European Semantic Web Conference ESWC 2020, Springer, 2020.

[6] V. W. Anelli, T. D. Noia, D. Malitesta, F. A. Merra, Assessing perceptual and recommendation mutation of adversarially-poisoned visual recommenders (short paper), in: DP@AI*IA, volume 2776 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 49–56.

[7] T. D. Noia, D. Malitesta, F. A. Merra, Taamr: Targeted adversarial attack against multimedia recommender systems, in: DSN Workshops, IEEE, 2020, pp. 1–8.

[8] V. W. Anelli, T. D. Noia, F. A. Merra, The idiosyncratic effects of adversarial training on bias in personalized recommendation learning, in: RecSys 2021: Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands, ACM, 2021. URL: https://doi.org/10.1145/3460231.3478858. doi:10.1145/3460231.3478858.

[9] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, Recommender Systems - An Introduction, Cambridge University Press, 2010. URL: http://www.cambridge.org/de/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/recommender-systems-introduction?format=HB.

[10] S. Alonso, J. Bobadilla, F. Ortega, R. Moya, Robust model-based reliability approach to tackle shilling attacks in collaborative filtering recommender systems, IEEE Access 7 (2019) 41782–41798.

[11] M. P. O'Mahony, Towards robust and efficient automated collaborative filtering, Ph.D. thesis, Citeseer, 2004.

[12] S. K. Lam, J. Riedl, Shilling recommender systems for fun and profit, in: WWW, ACM, 2004, pp. 393–402.
[13] R. D. Burke, B. Mobasher, R. Bhaumik, C. Williams, Segment-based injection attacks against collaborative filtering recommender systems, in: ICDM, IEEE Computer Society, 2005, pp. 577–580.
[14] W. Zhou, J. Wen, Q. Xiong, M. Gao, J. Zeng, SVM-TIA a shilling attack detection method based on SVM and target item analysis in recommender systems, Neurocomputing 210 (2016) 197–205.
[15] M. Aktukmak, Y. Yilmaz, I. Uysal, Quick and accurate attack detection in recommender systems through user attributes, in: RecSys, ACM, 2019, pp. 348–352.
[16] F. Zhang, Y. Lu, J. Chen, S. Liu, Z. Ling, Robust collaborative filtering based on non-negative matrix factorization and r_1-norm, Knowl.-Based Syst. 118 (2017) 177–190.
[17] G. Adomavicius, J. Zhang, Impact of data characteristics on recommender systems performance, ACM Trans. Management Inf. Syst. 3 (2012) 3:1–3:17.
[18] Y. Deldjoo, T. D. Noia, E. D. Sciascio, F. A. Merra, How dataset characteristics affect the robustness of collaborative recommendation models, in: SIGIR, ACM, 2020, pp. 951–960.
[19] B. Mobasher, R. Burke, R. Bhaumik, C. Williams, Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness, ACM Transactions on Internet Technology (TOIT) 7 (2007).
[20] M. P. O'Mahony, N. J. Hurley, G. C. M. Silvestre, Recommender systems: Attack types and strategies, in: AAAI, 2005, pp. 334–339.
[21] M. P. O'Mahony, N. J. Hurley, G. C. Silvestre, An evaluation of the performance of collaborative filtering, in: 14th Irish Artificial Intelligence and Cognitive Science (AICS 2003) Conference, Citeseer, 2003.
[22] M. P. O'Mahony, N. J. Hurley, N. Kushmerick, G. C. M. Silvestre, Collaborative recommendation: A robustness analysis, ACM Trans. Internet Techn. 4 (2004) 344–377.
[23] Y. Moshfeghi, B. Piwowarski, J. M. Jose, Handling data sparsity in collaborative filtering using emotion and semantic based features, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 625–634.
[24] P. Cremonesi, A. Tripodi, R. Turrin, Cross-domain recommender systems, in: ICDM Workshops, IEEE Computer Society, 2011, pp. 496–503.
[25] Y. Deldjoo, T. D. Noia, F. A. Merra, Assessing the impact of a user-item collaborative attack on class of users, in: ImpactRS@RecSys, volume 2462, CEUR-WS.org, 2019.
[26] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: CSCW 2000, Philadelphia, PA, USA, December 2-6, 2000, 2000, pp. 241–250. URL: https://doi.org/10.1145/358916.358995. doi:10.1145/358916.358995.
[27] C. C. Aggarwal, Recommender Systems - The Textbook, Springer, 2016. URL: https://doi.org/10.1007/978-3-319-29659-3. doi:10.1007/978-3-319-29659-3.
[28] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, TiiS 5 (2016) 19:1–19:19.
[29] X. He, Z. He, X. Du, T. Chua, Adversarial personalized ranking for recommendation, in: SIGIR, ACM, 2018, pp. 355–364.
[30] M. Schedl, The lfm-1b dataset for music retrieval and recommendation, in: ICMR, ACM,

2016, pp. 103–110.
[31] N. Hug, Surprise, a Python library for recommender systems, 2017.
[32] Y. Koren, Factor in the neighbors: Scalable and accurate collaborative filtering, TKDD 4 (2010) 1:1–1:24.
[33] Y. Koren, R. M. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, IEEE Computer 42 (2009) 30–37.
[34] P. Cremonesi, Y. Koren, R. Turrin, Performance of recommender algorithms on top-n recommendation tasks, in: Proc. of the 2010 ACM Conference on Recommender Systems, RecSys 2010, 2010, pp. 39–46. URL: https://doi.org/10.1145/1864708.1864721. doi:10.1145/1864708.1864721.