Who is Really Affected by Fraudulent Reviews?
An analysis of shilling attacks on recommender systems in real-world scenarios

Anu Shrestha
Boise State University
anushrestha@u.boisestate.edu

Francesca Spezzano
Boise State University
francescaspezzano@boisestate.edu

Maria Soledad Pera
Boise State University
solepera@boisestate.edu

ABSTRACT
We present the results of an initial analysis conducted to understand the effect of fake reviews or fake ratings in a real-world setting. We focus on both algorithm performance as well as the types of users who are most affected by these attacks.

KEYWORDS
Shilling attacks; spam reviews; reliable users

1 INTRODUCTION
The effect of shilling attacks on recommender systems, where malicious users create fake profiles so that they can then manipulate algorithms by providing fake reviews or ratings, has been long studied. Previous work has characterized and modeled shilling attacks on recommenders, defined new metrics to quantify the impact of these attacks on known recommender algorithms, and applied a detect-filtering approach to mitigate the effects of spammers on recommendations (see recent survey [1]). We observe from the literature that empirical analysis thus far has focused on assessing the robustness of recommender systems via simulated attacks [1, 10]. Unfortunately, there is lack of evidence on what is the impact of fake reviews or fake ratings in a real-world setting.

We present a preliminary analysis conducted to understand the influence of fraudulent reviews on the recommendation process. We do this through an initial study on known datasets with gold standards in different domains and a commonly-used recommendation algorithm. Our goal is to shed light on the effect of this attack and identify gaps to be addressed in the future by seamlessly connecting recommender and data mining research, as the latter has a rich body of work when it comes to spam detection and prevention.

2 ANALYSIS FRAMEWORK
Datasets. We use two real-world datasets (Table 1) that offer information about fraudulent reviews, which we treat as ground truth.

| Dataset       | Users | Items | Ratings | Spammers |
|---------------|-------|-------|---------|----------|
| Amazon-Beauty | 167,725 | 29,004 | 252,056 | 3.26%    |
| Amazon-Health | 311,636 | 39,539 | 428,781 | 4.12%    |
| Yelp!-Hotel   | 5,027   | 72     | 5,857   | 14.92%   |
| Yelp!-Restaurant | 34,523 | 129   | 66,060  | 20.25%   |

Table 1: Summary of datasets

Yelp! [8]: This dataset consists of Yelp reviews from two domains: hotels (YH) and restaurants (YR). Yelp filters fake/suspicious reviews and puts them in a spam list. A study found the Yelp filter to be highly accurate [11] and researchers have used filtered spam reviews as ground truth for spammer detection (e.g., [9]). Spammers, in our case, are users who wrote at least one filtered review.

Amazon [6]: Here we consider reviews from two domains: beauty (AB) and health (AH). In this case, we define ground truth following the framework in [2], which is based on helpfulness votes. Thus, we treat as a spammer every user who wrote at least one review in which he rated a product as 4 or 5 and has helpfulness ratio ≤ 0.4.

Experimental setting. In this paper, we analyze the robustness to shilling attacks of matrix factorization (MF), a commonly-used recommender algorithm. We used probabilistic MF [7] with 40 latent factors and 150 iterations. We performed 5-fold cross-validation and measured the performance in terms of RMSE only for non-spam users. We also used prediction shift (PS), a measure explicitly defined to quantify the impact of spammer attacks on recommenders, which captures the average changes in predicted ratings [1] ².

3 RESULTS & DISCUSSIONS
We discuss the effect of fake reviews on recommendations offered to users in real-world scenarios, as opposed to simulated attacks.

Do spam ratings affect recommendations? By following the classical evaluation framework for shilling attacks on recommender systems [1], we measured the performances on the original dataset (with spam) and when we remove all the reviews written by spammers (shilling attack). We report the results of our assessment in Table 2.

We anticipated a lower RMSE when removing spam. However, we did not observe this trend among most datasets in our study. This result aligns with previous work reporting (simulated) shilling attacks are not detectable using traditional measures of algorithm performance [5]. Previous works also show PS values ranging from 0.5 to 1.5 when shilling attacks are simulated. However, we observe very low values in real-world scenarios: in our case, considered, PS ranges from 0.047 to 0.15, which we argue is not enough to promote

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored.

LRRS@RecSys'18, October 2-7, 2018, Vancouver, BC, Canada
© 2018 Copyright held by the owner/authors.

1 Using hitRatio, we obtained similar outcomes. Thus, due to space limitations, we excluded that metric from our discussion.

2 In addition to PS, we considered stability of prediction, another common measure to quantify spammer attacks. As it is inversely proportional to PS, we only report PS.
or demote products attacked by the spammers. We believe this to be one of the reasons why algorithm robustness is not reflected by average metrics like RMSE. Further, looking at users as a whole does not help us quantify how much spammers are able to deceit recommenders or who are the users that are affected the most.

Table 2: RSME and PS on datasets with and without spam.

| Dataset            | W/ Spammers (RMSE, PS) | W/o Spammers RMSE |
|--------------------|------------------------|-------------------|
| Amazon-Beauty      | (0.871, 0.122)         | 0.901             |
| Amazon-Health      | (1.056, 0.047)         | 1.053             |
| Yelp!-Hotel        | (1.124, 0.150)         | 1.125             |
| Yelp!-Restaurant   | (1.039, 0.133)         | 1.034             |

4 CONCLUSIONS & FUTURE WORK

We have presented the results of an initial empirical analysis that has allowed us to demonstrate that trends observed as a result of simulated shilling attacks on recommender algorithms remain the same in a real-world scenario. We validated that average metrics are not able to properly capture attack effect and that in the presence of spam, recommender algorithms are not uniformly robust for all type of benign users. These initial discoveries lead us to argue in favor of new algorithms that are not only robust to attacks, but that also ensure that all users are protected against spam while supporting spam detection that accurately spots the subset of spammers who in fact affect recommendations without mistreating non-traditional users (i.e., users whose taste differs from the popular) as spammers.

REFERENCES

[1] Robin Burke, Michael O’Mahoney, and Neil Hurley. 2015. Robust collaborative recommendation. In Recommender systems handbook. Springer, 961–995.
[2] Amir Fayazi, Kyunum Lee, James Caverlee, and Anna Squicciarini. 2015. Uncovering Crowdsourced Manipulation of Online Reviews. In SIGIR. 233–242.
[3] Jennifer Golbeck. 2006. Generating predictive movie recommendations from trust in social networks. In Trust. Springer, 93–104.
[4] Srijan Kumar, Francesca Spezzano, V. S. Subrahmanian, and Christos Faloutsos. 2016. Edge Weight Prediction in Weighted Signed Networks. In ICDM. 221–230.
[5] Shyong K. Lam and John Riedl. 2004. Shilling recommender systems for fun and profit. In WWW. 393–402.
[6] Julian McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In WWW. 897–908.
[7] Andriy Mnih and Ruslan Salakhutdinov. 2008. Probabilistic matrix factorization. In NIPS. 1257–1264.
[8] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie S. Glance. 2013. What Yelp Fake Review Filter Might Be Doing?. In JCMW. 409–418.
[9] Shebani Rayana and Leman Akoglu. 2015. Collective Opinion Spam Detection: Bridging Review Networks and Metadata. In SIGKDD. 985–994.
[10] Carlos E Segui. 2013. Accuracy and robustness impacts of power user attacks on collaborative recommender systems. In RecSys. ACM, 447–450.
[11] Karen Weise. 2011. A Lie Detector Test for Online Reviewers. In Bloomberg BusinessWeek.