"Reverse Attack: Black-box Attacks on Collaborative Recommendation" Summary

Robert Krzysztof Robert Noparlik

November 4, 2023

1 Summary

This paper presents an approach to maliciously influence the recommendation systems of websites for the attacker's benefit. This approach works by first gathering data on the target recommender system by noting which recommendations appear for which item. This data will be used to train a "surrogate model" meant to replicate the real world one. After training, this model is used for evaluating how many products and new user accounts should be created to influence the recommender system to deliver users to our target products. Experimental results of this paper suggest that, at the time, this approach fares better than any of the current methods in the existing literature, despite the fact that all of them require some knowledge about the type of the target recommender system.

2 Pros

- This approach does not require any knowledge about the targeted system. This greatly simplifies a potential attack.
- According to results of the conducted experiments, this approach seems to be the best of its class right now.

3 Cons

- Does not take into account content personalization.
- The authors did not open source their implementation (despite describing the "novel part" pretty thoroughly).

4 Meaning

The authors of this paper created a new approach to influence recommender systems. Unlike most literature, it deals with recommender systems as black boxes, only training the system on data in the quantity a normal user is able to tackle. This is a much more realistic scenario, but it greatly complicates the solution. Despite this, their approach seems to beat other, white-box based systems in this regard, which makes it even more surprising.However, one thing the paper lacks is per-user customization, which is prevalent in most modern websites.

It's hard to gauge the severity of recommender attacks. They are not outright illegal, but they do provide a way to exploit the system, which is, at the very least, unethical. Things like this are probably more critical on websites that base much of their existence on recommender systems, like YouTube for example. Being recommended allows content creators to make money off of the platform. This seems like an ideal use case for the methods described in the paper, as a potential attacker could simply lead users to his content and make money off of it. However, it is important to note, that in case of large platforms, the amount of work required to compete with user-generated outweigh potential gains of abusing the system.

5 Discussion

- As mentioned before, it seems like platforms which do not group content (ex. YouTube) and are large enough to be worthy of exploitation require an impossibly large amount of effort to game the system. An attacker would need to create a lot of users and collectively perform a lot of actions to potentially align the recommender system to his will.
- Based on this, my question is: when does it become unfeasible to deploy such an attack?