Optimal Attacking Strategy Against Online Reputation Systems with Consideration of the Message-based Persuasion Phenomenon

#3932
IEEE ICASSP 2021

Zhanjiang Chen and H.Vicky Zhao

Dept. of Automation, BRNist, Tsinghua University, Beijing, P.R.China

Background

Reputation Systems

Biased Ratings

https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/

Banerjee, 1992
Andreoni J, 2009
Zhang X, 2017
Xie H, 2019

Herding Behavior

Whitby A, 2004
Liu Y, 2013
Singh P, 2015
Fang M, 2020

Misbehavior Attacks

We study the impact of herd behavior on the attacking strategies with limited resources

Model

Reputation System

Each user gives a numerical rating between -1 and 1 for the target item in turn, where -1 means negative and 1 means positive

Rating Behavior Model

Following the work in [1], user ratings are often closely correlated with historical rating, as the existence of social psychological phenomena such as herding.

$$R_i = \gamma \widetilde{R}_i + (1-\gamma)\widehat{R}_i$$

Rating Given by the i-th User Social-impact Rating Objective Rating

Social-Impact Rating

$$s_i = \sum_{k=0}^{i} \eta^k R_{i-k} = \eta s_{i-1} + R_i$$

$$\widetilde{R}_i = \frac{1-\eta}{1-\eta^i} s_{i-1} \approx (1-\eta) s_{i-1}$$

The normalization term is to ensure that when all historical ratings are constant *r*, then social-impact rating should also be *r*.

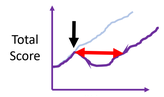
Strategy

Performance Metrics

We consider the scenario where attackers aim to downgrade ratings of their competitors and inject negative reviews.

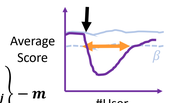
Total Reputation Score

$$LU = \inf\left\{i > m: \sum_{j=1}^{i} R_j \geq \sum_{j=1}^{m} R_j\right\} - m$$



Average Reputation Score

$$RU@\beta = \inf\left\{i > m: \frac{1}{i}\sum_{j=1}^{i} R_j \geq \beta \cdot \frac{1}{m}\sum_{j=1}^{m} R_j\right\} - m$$



Clicks

$$UT@\tau = \text{count}\ \{i: t_i \leq \tau\}$$

Theorem 1 Assume that attackers inject negative comments of length *L* at one time after *m* normal users submit their feedback, and the attack is strong enough, then we have

$$LU \approx \frac{E+1}{E}\left(L + \frac{\gamma}{1-\alpha}\right) \qquad RU@\beta \approx \frac{1}{1-\beta}LU - m$$

where $\alpha = \gamma + \eta - \gamma\eta < 1$.

Theorem 2 Assume attackers adopt the scheme and the attack is strong enough, then we have

$$LU \approx \frac{E+1}{E}\left\{CL + \frac{\gamma(1-\eta^L)}{1-\alpha} \cdot \left(\frac{\theta}{(1-\omega^2)} + \frac{1-\omega^C}{1-\omega}\right)\right\}$$

$$RU@\beta \approx \frac{1}{1-\beta}LU - m$$

where $\theta = (1-\alpha^I)\left((C-1)(1-\omega) - (\omega - \omega^C)\right)$, $\omega = \eta^{L+1} + \gamma(1-\eta)\eta^L\lambda$, $\lambda = (\eta^L - \alpha^L)/(\eta - \alpha)$. [2]

The Single-wave Attack

Attackers inject a single and continuous wave of *L* fake ratings.

The Multi-wave Attack

Attackers divide the fake ratings into several batches and launch several waves of attacks.

Optimal Strategy

It is difficult to find the close-solution of the optimal attacking strategy.

$$\max_{C,L,I} \quad LU \quad (RU@\beta \text{ or } -UT@\tau)$$

$$\text{s.t.} \quad C \cdot L = C_0$$

$$C, L, I \in \mathbb{Z}^+$$

Simulation

The Single-wave Attack

Attackers should inject more fake ratings as soon as possible to maximize the impact of the attack.

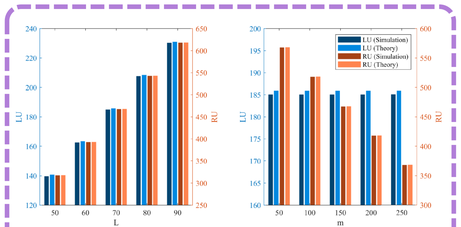


Fig. 1. LU and RU@70% under different single-wave attacks

The Multi-wave Attack

Set the total number of injected fake ratings to *C*·*L* = 60 and analyze the changes of the optimal attack strategy in different settings.

When users are more rational, attackers need to inject more fake ratings into each wave of attacks.

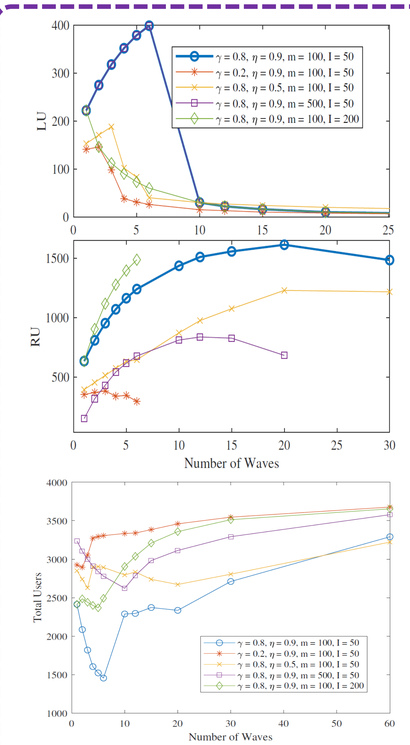


Fig. 2. LU, RU@70% UT@3000 under different multi-wave attacks

Conclusion

- We **jointly** model the attack on and the herd behavior in reputation systems.
- We propose a method to find the **optimal attack strategy** for a simple reputation system exploring the "message-based persuasion" phenomenon.
- We compare the **single-wave attack** with the **multi-wave attack** and find the optimal parameters that maximize the impact of the attack.

References

[1] Hong Xie, Yongkun Li, and John CS Lui, "Understanding persuasion cascades in online product rating systems," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, pp. 5490–5497.

[2] Zhanjiang Chen and H. Vicky Zhao, "Supplementary information of icassp 2021: Optimal attacking strategy against online reputation systems with consideration of the message-based persuasion phenomenon," [EB/OL], 2020, https://1drv.ms/b/s!ApDxURCRgQ9mgalypGGCA7foxMphOg.

[3] Yuhong Liu, Yan Sun, Siyuan Liu, and Alex C Kot, "Securing online reputation systems through trust modeling and temporal analysis," IEEE transactions on information forensics and security, vol. 8, no. 6, pp. 936–948, 2013.

[4] Hong Xie, Yongkun Li, and John CS Lui, "Optimizing discount and reputation trade-offs in e-commerce systems: Characterization and online learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, pp. 7992–7999.