JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ADOPT TO COMBAT MISINFORMATION ATTACKS IN OTHER REPUTATION SYSTEMS FOR RUMOUR DETECTION ON SOCIAL NETWORKS

Kesana Sai Gayathri Jnana Prasanna, A. Mary Sowjanya, G. Kumari

Department of Computer Science and Systems Engineering, Andhra University College of Engineering(A), Visakhapatnam, Andhra Pradesh

***Abstract*—** Online reviews provide customers with product evaluations to make decisions. Unfortunately, fake reviews ("spam") can be used to manipulate assessments by professional spammers, who have learned increasingly insidious and powerful spam tactics by adapting the detectors they deploy. Spam tactics are difficult to capture because they can change rapidly over time, vary between spammers and target products, and, crucially, remain unknown in most cases. Furthermore, most existing detectors focus on detection accuracy, which is inconsistent to maintain confidence in product evaluations. To address these challenges, we formulate a minimax game in which spammers and spam detectors compete on their actual goals, not just based on detection accuracy. The Nash equilibrium of the game leads to a stable detector that is not affected by any mixed detection strategy. However, the game has no closed-form solutions and cannot differentiate between typical gradient-based algorithms. We turn the game into two dependent Markov Decision Processes (MDPs) to allow efficient stochastic optimization based on multi-armed bandits and policy gradients. We conduct experiments on three large review datasets using various state-of-the-art spam and detection strategies and show that the optimization algorithm can reliably find a balanced detector that is robust and effective against the adoption of any mixed spam strategy of spammers to achieve their actual goals. Our code is available at https://github.com/YingtongDou/Nash-Detect.

Keywords— Spam detection, Nash-Detect, Accuracy, Minimax game, Markov Decision Process.

1. Introduction: Online reviews and ratings from real customers help shape a business' reputation and guide customer decisions, acting across e-commerce and sites like Amazon, Yelp, and Google Play. However, the monetary incentives included also attract a large number of spammers to control unwitting customers: it is estimated that about 40% of reviews on Amazon are fake (known as "review spam"). To deal with spam and restore the credibility of online reviews, many detection methods based on text reviewer behavior and graphs have been proposed. See Table 1 for some of the latest technologies. We note two shortcomings of existing detectors. 1) Most detectors assume that spam is generated by spammers with the same mindset, and can rely on hypothetical spam policies to detect spam. In the real world, there are multiple groups of spammers with different goals, objectives, resources, and strategies. One spammer might want to promote a new business, while another aims to demote a popular brand's competitor. The wide range of detection signals published to date strongly demonstrates that multiple spam strategies coexist and that no single detector can stop spam.2) Professional spammers are more persistent and focused, and can study the latest detection techniques from published papers, third-party detection sites with detailed detection manuals, and penetrate deployed detectors.

2. Literature Survey:

Adam Breuer et al. [1] proposed Graph-Based Early Detection of Fake Accounts on Social Networks. The problem of early detection of fake user accounts on social networks is based solely on their network connectivity with other users. Removing such accounts is

a core task for maintaining the integrity of social networks, and early detection helps to reduce the harm that such accounts inflict. However, new fake accounts are notoriously difficult to detect via graph-based algorithms, as their small number of connections is unlikely to reflect a significant structural difference from those of new real accounts. We present the Sybil Edge algorithm, which determines whether a new user is a fake account (`sybil') by aggregating over (I) her choices of friend request targets and (II) these targets' respective responses. Sybil Edge performs this aggregation by giving more weight to a user's choices of targets to the extent that these targets are preferred by other fakes versus real users, and also to the extent that these targets respond differently to fakes versus real users. We show that Sybil Edge rapidly detects new fake users at scale on the Facebook network and outperforms state-of-the-art algorithms. We also show that Sybil Edge is robust to label noise in the training data, to the different prevalence of fake accounts in the network, and to several different ways fakes can select targets for their friend requests. To our knowledge, this is the first time a graph-based algorithm has been shown to achieve high performance (AUC>0.9) on new users who have only sent a small number of friend requests.

Sebastien Bubeck et al. [2] presented Regret Analysis of Stochastic and No stochastic Multi-armed Bandit Problems. Multi-armed bandit problems are the most basic examples of sequential decision problems with an exploration-exploitation trade-off. This is the balance between staying with the option that gave the highest payoffs in the past and exploring new options that might give higher payoffs in the future. Although the study of bandit problems dates back to the Thirties, exploration-exploitation trade-offs arise in several modern applications, such as ad placement, website optimization, and packet routing. Mathematically, a multi-armed bandit is defined by the payoff process associated with each option. In this survey, we focus on two extreme cases in which the analysis of regret is particularly simple and elegant: payoffs and adversarial payoffs. Besides the basic set of finitely many actions, we also analyze some of the most important variants and extensions, such as the contextual bandit model.

Yizheng Chen et al. [3] forwarded Practical Attacks Against Graph-based Clustering. Graph modeling allows numerous security problems to be tackled in a general way, however, little work has been done to understand their ability to withstand adversarial attacks. We design and evaluate two novel graph attacks against a state-of-the-art network-level, graph-based detection system. Our work highlights areas in adversarial machine learning that have not yet been addressed, specifically: graph-based clustering techniques, and a global feature space where realistic attackers without perfect knowledge must be accounted for (by the defenders) to be practical. Even though less informed attackers can evade graph clustering at a low cost, we show that some applicable defenses are possible.

L.Akoglu et al. [4] proposed Opinion fraud detection in online reviews by network effects. User-generated online reviews can play a significant role in the success of retail products, hotels, restaurants, etc. However, review systems are often targeted by opinion spammers who seek to distort the perceived quality of a product by creating fraudulent reviews. We propose a fast and effective framework, FRAUDEAGLE, for spotting fraudsters and fake reviews in online review datasets. B. Biggio et al. [5] preferred Security evaluation of pattern classifiers under attack. These results show that security evaluation can provide a more complete understanding of the classifier's behavior in adversarial environments, and lead to better design choices.

W. Dai et al. [6] analyzed Optimal aggregation of consumer ratings: an application to yelp. Com. In this paper, they analyzed reviews from Yelp.com to derive optimal ratings for each restaurant. C. Forman et al. [7] proposed, Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets Using research on information processing as a foundation, we suggest that in the context of an online community, reviewer disclosure of identity-descriptive information used by consumers to supplement or replace product information when making purchase decisions and evaluating the helpfulness of online reviews.

C. Elkan et al. [8] proposed the foundations of cost- -g,' sensitive learning. In this paper, we studied how to make decisions based on a cost matrix, and achieve cost sensitivity by rebalancing. Bryan Hooi et al. [9] preferred FRAUD: Bounding Graph Fraud in the Face of Camouflage. Shuaijun Ge et al. [10] analyzed Securing Behavior-based Opinion Spam Detection. In this paper, we studied the plot of total/negative/positive spam posted, the successful rate of evasion, and average promotions in CMR and CAR.

3. Methodology:

Nash-Detect also exhibits great stability under various settings during training. Experiments show that Nash-Detect can find the best detector configurations that always have better defensive performance than the worst-case scenarios that spammers can synthesize. The deployed experimental results show that different accounts and products should be combined during training to guarantee the robustness of the resulting detector.

Attacks based on text generation and spam speed controls exist, but for simplicity, we do not consider such controls in this work. All of these previous attacks did not differentiate between elite and normal accounts and were not trained to maximize actual spam effects. We consider cases where elite accounts and regular accounts contribute differently to the actual spam effect.

This observation further confirms that the promotions are mainly from a small number of elite accounts that evade detection, even when many singleton spam emails are detected.

3.1 Procedure:

- Dynamic game between Spammer and Defender
- Practical Evaluation Metric
- Evolving Spamming Strategies
- Multiple detectors Ensemble

Figure 2(a): A vulnerable spam detection pipeline. Accuracy-based detectors can be misled to detect numerous insignificant new accounts, leaving behind the more manipulative elite spam. We define a zero-sum game to find a robust defender against unknown and evolving spamming strategies A(p).

Figure 2(b): The Practical Effect vs. Recall of individual detectors (shown in legend) against a mixed spamming strategy. The curve is obtained by sweeping the detection thresholds. For most detectors, the attack could attain high practical effects even with high detection recall scores.

Figure 2(c): For a fixed spam detector (Fraudar), a spammer can choose the best out of five attack strategies to maximize the practical effect.

Algorithm 1 Nash-Detect: Training a Robust Spam Detector

Input: all reviews $\mathcal{R}$, target items $\mathcal{V}_T$, pure attack strategies $[a_1, \ldots, a_K]$, pure spam detectors $[d_1, \ldots, d_L]$, initial spamming strategy $\mathrm{p}^{(0)} = [p_1, \ldots, p_K]$ and initial detection strategy $\mathrm{q}^{(0)} = [q_1, \ldots, q_L]$ to uniform distributions.
Output: a Nash equilibrium $(\mathrm{p}^*, \mathrm{q}^*)$.
repeat ▷ Go through the H episodes indexed by t
  Inference:
  $\mathcal{R}(\mathrm{p}^{(t)}) = \mathcal{R}$.
  for all $v \in \mathcal{V}_T$ **do** ▷ Post fake reviews
    Sample a_k using ϵ-greedy for v according to $\mathrm{p}^{(t)}$.
    Post spams to v using a_k.
  Remove spams in the top k reviews detected by $D(\mathrm{q}^{(t)})$.
  Compute $\mathrm{PE}(v, \mathcal{R}, \mathrm{p}^{(t)}, \mathrm{q}^{(t)})$ using Eq. (2) on $\mathcal{R}(\mathrm{p}^{(t)}, \mathrm{q}^{(t)})$.
  Learning:
  Compute $C_{\mathrm{FN}}(v, r)$ using Eq. (4) and $G(a_k)$ using Eq. (9).
  Update $\mathrm{p}^{(t)}$ to $\mathrm{p}^{(t+1)}$ using the gains $G(a_k)$.
  Update $\mathrm{q}^{(t)}$ to $\mathrm{q}^{(t+1)}$ by minimizing Eq. (11).
until $\mathcal{L}(\mathrm{q})$ converges

Fig 1: Algorithm for Nash-Detect: Training a Robust Spam Detector

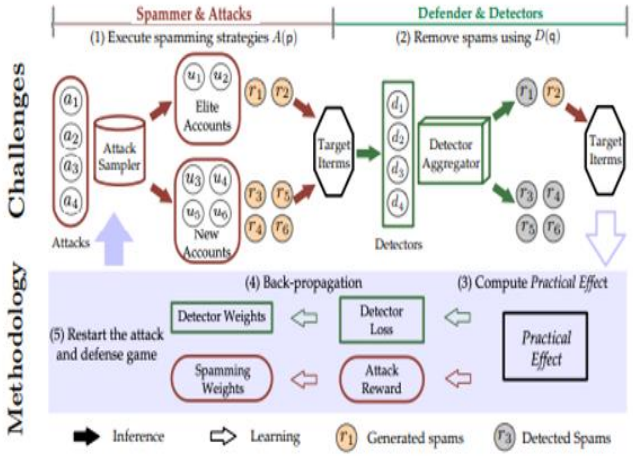

Fig 2(a): Challenges and methodology

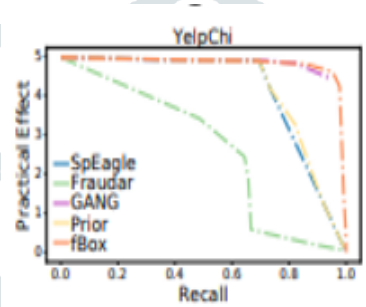

Fig 2(b): Practical effect vs. Recall for different detectors

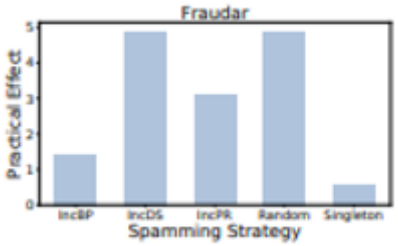

Fig 2(c): Practical effect under different spamming attack strategies

We present the following basic spam strategies, which differ in their target detectors and whether elite accounts are used, and let Nash-Detect understand the importance of each basic spam strategy.

3.2 Base Detection and Spamming Strategies:

As mentioned, there are many graph-based and behavior-based detectors. We select the following five base detectors:

• GANG: a social network Sybil detection algorithm via linearized belief propagation.

• SpEagle: an advanced belief propagation algorithm verified on Yelp review spam datasets.

• fBox: an SVD-based algorithm that spots small-scale suspicious links in social networks.

• Fraudar: a fraudster detection algorithm that detects dense blocks in graphs.

• Prior: an algorithm that ranks spam based on multiple suspicious behaviors.

These two strategies spamming detectors and reinforcement learning are expected to evolve into a balance consisting of robust detection strategies. To ensure computational tractability, we propose a reinforcement learning approach to find the balance of a minimax game across multiple fictional episodes of the game. Each episode has two phases. During the inference phase, the spammer samples the base spam policy according to the current hybrid policy, and the detector runs its current detection policy. This step will evaluate the actual spam performance under both current strategies.

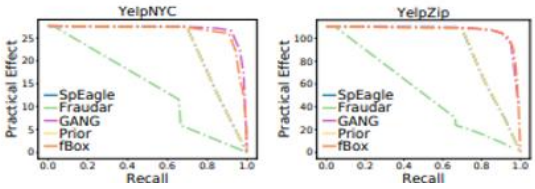

Fig 3: Practical Effect vs. Recall for different detectors against ensemble attacks on Yelp NYC and Yelp Zip.

Spam tactics are difficult to capture because they can change rapidly over time, vary between spammers and target products, and, crucially, remain unknown in most cases. Furthermore, most existing detectors focus on detection accuracy, which is inconsistent to maintain confidence in product evaluations. To address these challenges, we formulate a minimax game in which spammers and spam detectors compete on their actual goals, not just based on detection accuracy. The spammer's practical goal is to promote a product, and the practical goal of the spammer is to maximize the PE. The Defender Practical goal: The defender needs to minimize the practical effect. We combine the detector prediction results with the practical effect to formulate a cost-sensitive loss.

4. Results and Evaluation Metrics:

Algorithm Name	Review AUC	Review AP
GANG	0.7549622848885058	0.2879056981313653
fBox	0.49840065748649326	0.15451052213763217
Fraudar	0.2577786422820625	0.10945860689543202
Prior	0.677926167729431	0.2521225021708665
SpEagle	**0.7657989461836956**	**0.30167877459861897**

Table 1: Values of AUC and AP for five detectors and SpEagle is the most efficient spam detection algorithm

	GANG	SpEagle	fBox	Fraudar	Prior
IncBP	4.8916	4.9052	4.9125	1.4203	4.9099
IncDS	4.9010	4.9052	4.9110	4.8959	4.9099
IncPR	4.9010	4.9052	4.9105	3.0716	4.9099
Random	4.9010	4.9052	4.9092	4.8962	4.9099
Singleton	0.5300	0.5865	0.5783	0.5771	0.5912

Table 2: The practical effect of detectors against attacks under YelpChi

Table 2 shows the practical effect of individual detectors against individual attacks. One can see that if the spammer knows that a specific detector (such as Fraudar and fBox) is adopted by the defender, the spammer can adopt the spamming strategy (such as Random or IncBP) that leads to the most practical effect concerning the known detector. Therefore, a detector ensemble configuration D(q) is necessary.

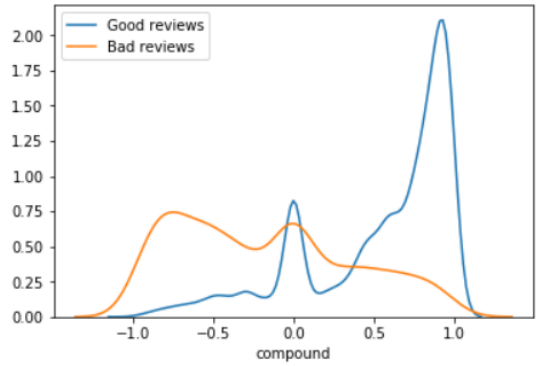

Fig 4: Graph shows the distribution of the review sentiments among good reviews and bad ones. We can see that good reviews are for most of them considered very positive. On the contrary, bad reviews tend to have lower compound sentiment scores.

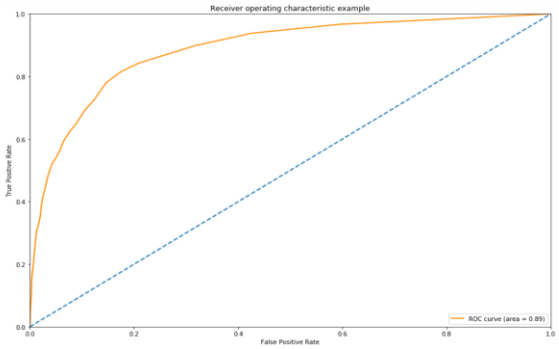

Fig 5: The ROC (Receiver Operating Characteristic) curve is usually a good graph to summarize the quality of our classifier. The higher the curve is above the diagonal baseline, the better the predictions.

5. Conclusion:

In this work, we propose a practical evaluation system metric that takes into account product revenue lift. We study the real-world performance of mainstream spam detectors against several target-oriented spam attacks in an adversarial setting. We formulate a game-theoretic model and reinforcement learning algorithm to find robust detection strategies against various spam strategies. Empirical evaluations on three large review datasets show that the proposed algorithm can indeed generate detectors that can effectively tame actual spam targets for product revenue manipulation. Our future work remains to adopt the proposed models and algorithms to combat misinformation attacks in other reputation systems, such as rumor detection on social networks. Inspired by the above work, we propose to incorporate economic incentives and the social status of accounts into the actual goals of spam. In a real-world business system, there are other factors, such as marketing operations and business categories that can also affect business revenue. However, it is not feasible to model such complex relationships using only online review data. Therefore, in this paper, we mainly focus on the relationship between ratings and revenue.

References:

[1] A. Breuer, R. Eilat, and U. Weinsberg. 2020. Friend or Faux: Graph-Based Early Detection of Fake Accounts on Social Networks. In WWW.

[2] S. Bubeck and N. Cesa-Bianchi. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. FTML (2012).

[3] Y. Chen, Y. Nadji, A. Kountouras, F. Monrose, R. Perdisci, M. Antonakakis, and N. Vasiloglou. 2017. Practical Attacks Against Graph-based Clustering. In CCS.

[4] L. Akoglu, R. Chandy, and C. Faloutsos. 2013. Opinion fraud detection in online reviews by network effects. In ICWSM.

[5] B. Biggio, G. Fumera, and F. Roli. 2013. Security evaluation of pattern classifiers under attack. In ICDE.

[6] W. Dai, G. Z. Jin, J. Lee, and M. Luca. 2012. Optimal aggregation of consumer ratings: an application to yelp. com. NBER WPS (2012).

[7] C. Forman, A. Ghose, and B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. ISR (2008).

[8] C. Elkan. 2001. The foundations of cost- -g,' sensitive learning. In IJCAI.

[9] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. 2016. FRAUD: Bounding Graph Fraud in the Face of Camouflage. In KDD.

[10] S. Ge, G. Ma, S. Xie, and P. S. Yu. 2018. Securing Behavior-based Opinion Spam Detection. In IEEE Big DATA.