Poster: Who spams whom? Detecting Opinion Spammer Groups and Their Spam Targets

Euijin Choo
Missouri University of Science and Technology
Email: chooe@mst.edu

***Abstract*—We investigate on detection of opinion spammer groups and their spam targets in review systems. People tend to trust reviews from top-ranked reviewers much more than those from low-ranked ones [1]. Spammers may thus artificially manipulate the ranking system to make their own reviews attract more attention from others; or to make competitors' reviews attract less attention. In this work, we thus address two aspects of potential for collaborative opinion spamming behavior (i.e., boosting or demoting). In our approach, we focus on outgoing relationships of spammers to detect users who are artificially promoted or demoted. Previously, we revealed strong positive spam communities based upon their interaction patterns and the sentiments of those interactions [2]. In this work we further explore positive/negative spam targets of such strong communities with the sentiment analysis on outgoing relationships from the strong communities. Through extensive experiments on Amazon dataset, we show that spammers tend to have interest more in their own promotion rather than in others' demotion.**

I. Introduction

There has been ample evidence that opinion spams are largely populated in practice that will eventually damage the service quality of review systems [3, 4]. There has been a growing body of research in opinion spam detection [5]. Although previous research has focused primarily on detecting positive spams using pure content-based classifiers, it is also critically needed to detect *spam targets* who pretend to be non-spammers but benefit from the collaborators' spams; *spam targets* who are demoted by their competitors [6]. In this work, we focus on spammers who spams targeted users through artificially orchestrated interactions.

II. Approach

Users' interactions in review systems without knowledge of each other are often assumed to occur randomly depending on their item interests [2]. If users have abnormal connections with others, on the other hand, reviews/replies by them can be biased, while favoring each other or disfavoring others . The goal of spammers would then be to promote their rankings through abnormal positive connections or to demote competitors' rankings through abnormal negative connections. To get promoted, spammers need significantly more positive responses than negative ones (e.g., helpful votes in Amazon). For the fairness, multiple votes from the same user on one review are often counted as one vote. Spammers therefore need to boost their ranks by gathering positive votes from different users (i.e., collusion). To do so, spammers may collaborate to express positive responses to each other. We thus hypothesize that such malicious artificial boosting activities would eventually lead to constructing communities in which spammers are strongly positively connected with each other through review-response interactions (e.g., votes and text replies on reviews). Similarly, spammers may collaborate to express negative responses to competitors. As spammers are unlikely to vote down for themselves but for competitors, we hypothesize that malicious demoting activities would eventually lead to strong negative connections from spammer groups to competitors. The goal of our research is thus to find these strongly or even abnormally positively connected communities and their positive/negative spam targets.

To define the abnormality of connections, we defined the strength of user connections based upon distance between a user's interaction pattern and a random model. User connections can be extended to communities so that a user belongs to τ strength of a community, if the user has τ strength of connections with another. The larger τ is, the stronger a community is. Connections that belong to stronger communities are excluded from weaker communities. That is, if a connection is in 99.5% community, it is excluded from 98% community. In our previous work [2], we have shown the correlation between the strength of connections and spammicity level. Specifically, we have observed users in communities whose strength is higher than 60%, tend to show activities deviating a lot from others in terms of spammicity level. We thus define those connections with strengths higher than 60% as abnormal connections.

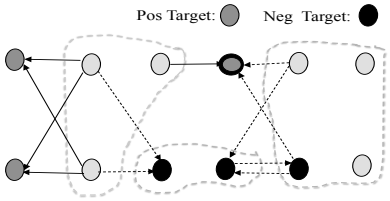

Fig. 1: Positive and negative targets of strong positive communities

To define the sentiments of connections, we compute the sentiment score ranging from -1 to 1 with a publicly available tool, AlchemyAPI [7], and aggregate the sentiments of all interactions between two users from which we derive the sentiment of each connection. That is, if their average sentiment score is more than 0, equal to 0, less than 0, we say they are having positive/neutral/negative connections, respectively.

To effectively spam targeted users, spammers first would need to make their opinions influential (i.e., become highly ranked users). In [2], we have shown that spammers are likely to build strong positive communities to achieve high ranks. We thus first find strong positive communities. Once we identify all strongly positively connected communities, we further analyze their outgoing relationships (i.e., relationships from the strong communities to others) to study their spam targets. To boost or demote others, spammers essentially need to form strong positive or negative relationships, not neutral relationships. We therefore build positive and negative outgoing relationship graphs of discovered communities by extracting positive and negative relationships from the strong communities to others, as illustrated in Fig.1.

III. Discussion

In [2], we collected reviews and replies across 4 item categories (Books, Movie, Electronics, Tools) from Amazon. Spammers may launch attacks not only in specific categories but also across categories. We thus also performed our experiments on the cross-category dataset, called Across. Among 5 dataset, we report results for the Across dataset, as the same patterns and observations were found for all individual categories.

Fig. 2 shows the number of 80% $\sim$ 99.5% of positive or negative relationships from strong positive communities to others. In [2], we have observed strongly positively connected communities with a strength higher than 80% are strong spammer candidates. We thus investigated outgoing relationships from those in the 80% $\sim$ 99.5% communities to others. We summarize our observation of 80% $\sim$ 99.5% strengths of outgoing relationships. First, there is a relatively large number of strong positive outgoing relationships, compared to the strong negative

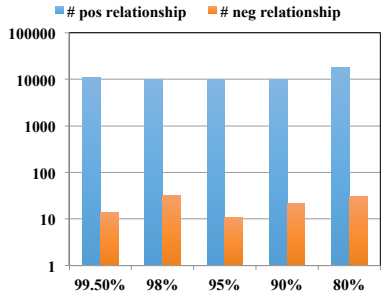

Fig. 2: The number of positive and negative relationships of spam reviewers in Across category

outgoing relationships. In general, strong positive outgoing relationships tended to appear inside the communities whose spammicity [2] was also high. Also, we have observed that there are not many demoting behavior in our dataset as shown in Fig. 2. This result might suggest that the primary goal of spammers is rather promoting their own reviews than demoting competitors'.

IV. Conclusion and Future Work

In this work, we aim to detect collaborative group spammers and their positive/negative spam targets. Towards this goal, we first detect strong positive communities as spammers, and analyze their outgoing relationships. Although a few researchers have suggested the critical impact of negative spams [3], our observation suggests that the targets of opinion spammers are often themselves, not the competitors. As for the future work, we plan to use the spam target information to build a spam alert system that issues spam alerts of being victims for spamming.

References

[1] M. Anderson, "Customer survey," 2013, http://searchengineland.com/2013-study-79-of- consumers-trust-online-reviews-as-much-as-personal- recommendations-164565.

[2] E. Choo, T. Yu, and M. Chi, "Detecting opinion spammer groups through community discovery and sentiment analysis," in *Data and Applications Security and Privacy XXIX*. Springer, 2015, pp. 170–187.

[3] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam." in *HLT-NAACL*, 2013, pp. 497–501.

[4] BBC, "Yelp admits a quarter of submitted reviews could be fake," 2013, http://www.bbc.com/news/technology-24299742.

[5] M. Jiang, P. Cui, and C. Faloutsos, "Suspicious behavior detection: Current trends and future directions," *Intelligent Systems, IEEE*, vol. 31, no. 1, pp. 31–39, 2016.

[6] Y. Liu, Y. L. Sun, and T. Yu, "Defending multiple-user-multiple-target attacks in online reputation systems," in *Privacy, security, risk and trust (passat), 2011 ieee 3rd int'l conf. on and 2011 ieee 3rd int'l Conf. on social computing (socialcom)*. IEEE, 2011, pp. 425–434.

[7] AlchemyAPI, publicly available at http://www.alchemyapi.com/.