Poster: Toward Understanding Spamming Behavior in Public Forums

Euijin Choo
North Carolina State University
Email: echoo@ncsu.edu

Ting Yu
North Carolina State University
Qatar Computing Research Institute
Email: tyu@ncsu.edu; tyu@qf.org.qa

Min Chi
North Carolina State University
Email: mchi@ncsu.edu

***Abstract*—Users often express their opinions about an item in which they are interested in public forums. Despite the convenience, it also opens possibilities for attackers to manipulate such forums with opinion spams. Recently, researchers have begun to pay attention to opinion spams. While previous research has focused primarily on detection of opinion spams based on text contents, relatively little is done to capture spammers' behavior based on user relationships. In this work we study opinion spam phenomena by detecting malicious communities in public forums. Concretely, we focus on review systems (e.g., Amazon, Yelp). In our previous work we found there exist users who have strong connections to each other in Amazon based on their replying actions, which is not desirable because their opinions are likely to be biased towards each other. We show that such users indeed tend to show spamming behavior by applying spam indicators to the users.**

I. Introduction

User-generated opinions increasingly play a key role in people's decision making process, which makes it easy for attackers to exploit with opinion spams [1, 2]. A few methods to detect opinion spams have been proposed in the literature [3–5], most of which focus primarily on detection of spams by developing supervised classifiers that compare text contents of unlabelled data and ground truth dataset. In this work, instead, we attempt to provide a glimpse of a review system to understand users' spamming behavior in terms of social connections between users in a system. Concretely, we investigate whether the fact that users deliberately reply to each other's reviews could give us any hint of spamming behavior.

Our work is grounded in the context of a review ecosystem in Amazon. In our previous work [6] we found that there exist users who form strong connections to each other through reviews/replying activities, which can in turn be extended to strong communities. We also found that such users tend to show activities deviating a lot from others. In this work we measure correlation between such communities and malicious activities (i.e., opinion spams). We show that users who have strong connections show a similar behavior pattern to spammers and thus such users can potentially be strong candidates for spammers.

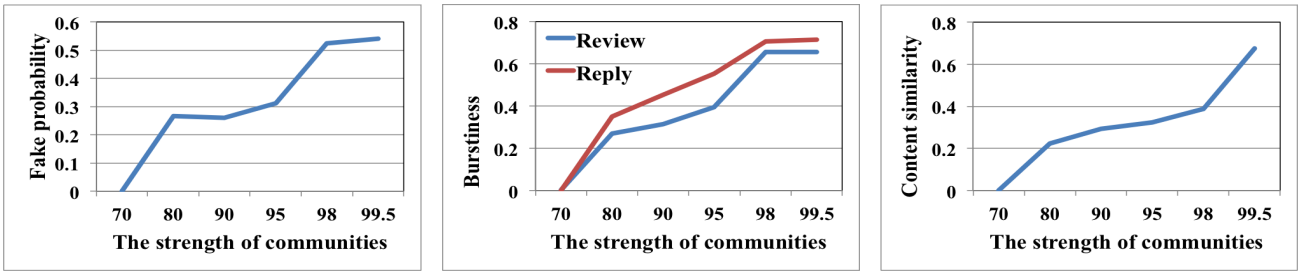
II. Spammicity of Communities

Users' replying actions in review systems without knowledge of each other (e.g., they have social connections with each other) are often assumed to happen by chance [6]. Users will frequently browse for an item of interest, read a review, and post a simple reply without considering the author of the review. If users have connections with each other, on the other hand, reviews/replies by them can be biased, favoring each other, which, in turn, results in unfair systems. We thus attempt to measure whether communities consisting of users having strong connections with each other are relevant to opinion spams in this section.

In [6], we collected reviews and replies across 4 item categories (Books, Movie, Electronics, Tools) from Amazon. We defined the strength of user connections based upon distance between a user's replying pattern in the collected dataset 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.

We summarize our observation of three spam indicators on the communities found in Amazon. We employed three spam indicators introduced in past work including fake probability [4], burstiness [7], and content similarity [7] as follows. Each value ranges from 0 (non-spammers) to 1 (spammers).

Fake probability estimates likelihood of each community user's reviews/replies being spams based on



(a) Fake probability (b) Burstiness (c) Content similarity

Fig. 1: Spammicity of communities with different strength

linguistic features.

Burstiness measures an interval between each community user's first review/reply and last review/reply.

Content similarity measures how similar a user's reviews/replies are to each other.

Among the 4 item categories we collected, we present results in the Movie dataset because the same results hold for the other three categories. Fig.1 shows three spammicity indicators (Y-axis) of communities discovered in the Movie dataset with different strength (X-axis). We observe that spammicity increases as the strength of communities increases, as shown in Fig.1. Note that spammicity drops dramatically between 70% and 80% communities. This suggests that such weaker connections (i.e., smaller than 80%) are often formed naturally (e.g., common interest in items). That is, users in weaker communities are not likely to be spammers as spammicty of users in 70% community was almost close to 0; whereas spammicity of users in stronger communities such as 99.5 was relatively high. Specifically, three spammicity indicators of such users were higher than 0.5, which is the threshold used to define spams in [4, 7], on average.

III. Conclusion and Future Work

Attackers may post fraudulent positive/negative reviews/replies to their colluders' or competitors' to boost or decrease the reputation of specific reviews or reviewers. Among such attackers, some users even may not have any reviewing history yet, but only engage in fraudulent replies to avoid detection by spam classifiers based on review contents. Our observation suggests that such malicious repliers in strong communities are strong candidates for future spammers; whereas existing supervised classifiers may not be able to find such users because their detection often bases on existing reviews.

Some challenges still must be surmounted. First, some users may naturally form a community because of their common interest on items, while the strong user connections were shown to be relevant to spamming behavior. It is thus needed to derive a method narrowing down the definition of spammers to differentiate natural and malicious communities and to reduce false positive of detections. Further, there may be more spam indicators other than three, so we need to build more indicators. Especially, we will derive not only spam review indicators but also spam replying indicators. Finally, we will derive a method to prevent spammers without previous spams who are most likely to post them in the future. Our experiment results actually suggest a way to prevent such spammers since we find spammer candidates based solely on users' replying patterns not on users' past spam contents.

References

[1] M. Hines, "Scammers gaming youtube ratings for profit," publicly available at http://www.infoworld.com/d/security-central/scammers-gamingyoutube-ratings-profit-139.

[2] C. N. Dellarocas, "Strategic manipulation of internet opinion forums: Implications for consumers and firms," pp. 460–473, 2004, publicly available at SSRN: http://ssrn.com/abstract=585404.

[3] Y. Liu and Y. L. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Proc. of IEEE International Conf. on Social Computing*, 2010, pp. 65–72.

[4] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 309–319.

[5] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proc. of the 21st international conf. on World Wide Web*, 2012, pp. 191–200.

[6] E. Choo, T. Yu, M. Chi, and Y. Sun, "Revealing and incorporating communities to improve recommender systems," in *Proc. of the 15th ACM conf. on economics and computation*, 2014.

[7] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," in *Proc. of the 19th ACM SIGKDD international conf. on Knowledge discovery and data mining*, 2013, pp. 632–640.