Detecting Opinion Spammer Groups Through Community Discovery and Sentiment Analysis

Euijin Choo, Ting Yu, Min Chi

Detecting Opinion Spammer Groups through Community Discovery and Sentiment Analysis

Euijin Choo[1], Ting Yu[1,2], and Min Chi[1]

[1] North Carolina State University
echoo,tyu,mchi@ncsu.edu,
[2] Qatar Computing Research Institute
tyu@qf.org.qa

Abstract. In this paper we investigate on detection of opinion spammer groups in review systems. Most existing approaches typically build pure *content-based* classifiers, using various features extracted from review contents; however, spammers can superficially alter their review contents to avoid detections. In our approach, we focus on user relationships built through interactions to identify spammers. Previously, we revealed the existence of implicit communities among users based upon their interaction patterns [3]. In this work we further explore the community structures to distinguish spam communities from non-spam ones with sentiment analysis on user interactions. Through extensive experiments over a dataset collected from Amazon, we found that the discovered strong positive communities are more likely to be opinion spammer groups. In fact, our results show that our approach is comparable to the existing state-of-art content-based classifier, meaning that our approach can identify spammer groups reliably even if spammers alter their contents.

Keywords: Opinion spammer groups, sentiment analysis, community discovery

1 Introduction

There has been a rapid and growing interest in recent years in opinion spamming [8–10, 13, 15, 17, 19]. Opinion spamming refers to malicious activities that aim to influence normal users' decisionmaking for profit.

While a number of methods have been proposed to detect opinion spam, most of them focus primarily on developing pure *content-based* classifiers [4, 10, 13, 17, 19]. The basic idea behind these approaches is to detect opinion spam through the analysis of review content. Such pure content-based classifiers, however, are limited for several reasons. First, spammers can easily manipulate review content to avoid detection [10, 17]. For example, if duplicated text reviews are considered to be spam, spammers may simply paraphrase the content. Second, they are often designed for specific application domains such as travel reviews, and cannot be applied easily to different domains such as movie reviews [13]. Third, while most content-based classifiers generally require ground truth labels, it is often hard to obtain for real datasets. Some previous researchers have hired human experts to manually label data. The high cost of this approach, however, makes it impossible to do so reliably for large-scale datasets [9].

In this paper we explore an alternative approach by examining what we call *promotional opinion spammers* through the analysis of user relationships rather than review content. Promotional opinion spammers refer to attackers who try to improve the influence of their opinions by malicious artificial boosting. For example, many review systems employ some sort of reviewer/review ranking systems e.g., a top reviewer list on Amazon, most helpful reviews on Amazon, or most recommended reviews on Yelp. Spammers may thus artificially boost the rank of their reviews to attract more attention.

To obtain high ranking, spammers need to collect significantly more positive responses than negative ones. For example, review and reviewer ranks on Amazon are based primarily on the number of *helpful* votes received. Since multiple votes from the same user on one review are often counted as one vote, spammers need to boost their ranks by gathering positive votes from different users (i.e., collusion). One possible way to do this is for spammers to collaborate to vote high. We thus hypothesize that such malicious artificial boosting activities would eventually lead to construct spammer communities in which spammers are strongly positively connected with each other through review-response interactions (e.g., votes and text replies on the review). Our goal is thus to find these strongly or even abnormally positively connected communities among users and we argue that it is more likely to detect collusive spamming behavior among these users than those who are not part of these communities.

Our work is grounded in the context of a review ecosystem on Amazon. In our prior work we identified the existence of implicit communities built through review/response activities on Amazon [3]. In this paper we further explore positively and negatively connected communities through review and response activities via sentiment analysis. The intuition behind this approach is that: if a user has an unusual positive or negative relationship with another, they may be posting fraudulent positive and negative responses to each other's items and/or reviews to boost or demote the reputation of specific reviews or reviewers. In this paper we focus on spammers' boosting behavior.

In our approach, we first build general user relationship graphs representing how users *interact* with one-another. Then, we derive the sentiment of each relationship by aggregating sentiments of all responses between any two users. We then extract *positive* relationship graphs from the general user relationship graphs to capture boosting behavior. More specifically, motivated by link-based web spam detection, we focus on strongly connected communities in positive relationship graphs. Finally, we analyze extracted strongly positively connected communities to find opinion spammer groups.

Note that non-spammers may also form natural communities based upon their genuine similar interests [3]. However, we argue that spammer communities have distinguishing characteristics in terms of structures and the strength of their relationships. Concretely, we show that the stronger a community the user appears in, the more likely the user is involved in spamming-like activities.

Our main contributions are summarized as follows.

(1) We propose a general unsupervised hybrid approach that is based on user interactions coupled with sentiment analysis. To the best of our knowledge, this is the first attempt to identify opinion spammer groups through analyzing users' interactions rather than their review content. A key advantage of our approach is that it can detect opinion spammers even when traditional review content-based approaches fail.

(2) We introduce a new angle of collusive spamming behavior that spammers deliberately build strong positive communities to make their own opinions influential. We thus propose to explore community structures and a strength of relationships (i.e., how much the relationships are likely to be built intentionally) as spam indicators.

(3) We run extensive experiments over a dataset collected from Amazon to evaluate the effectiveness of the proposed approach. Our experiments show that even though our community-based approach differs markedly from pure content-based approaches, it reaches the same level of accuracy as the state-of-art content-based approaches.

The remaining parts of this paper are organized as follows. In Section 2, we review related work. Section 3 presents an abstract model of review systems, and introduces basic concepts and notations used through this paper. Section 4 offers the proposed approach to analyze and detect spammer groups. In Section 5, we discuss experimental results. Finally, Section 6 concludes the paper.

2 Related Work

Unlike traditional spam analysis in the context of Web and email, it is often hard to get ground truth for opinion spam. Previous research employed different mechanisms to obtain ground truth data. Early work manually inspected reviews and extracted simple features such as duplicated reviews or unexpected rating patterns [4–7]. These approaches were limited as they depend largely on heuristics [9, 10].

A few researchers have created ground truth data by hiring Turkers to write spams [12, 13]. They then developed content-based classifiers that compare the linguistic features of genuine and spam reviews. While these classifiers have been shown to be successful, it is questionable whether they can be applied in other domains as they are content specific. For example, linguistic features of hotel and book reviews may be different. More importantly, there have been unresolved debates on whether datasets generated by Turkers can be representative of actual spams in practice [9].

Mukherjee *et al.* generated ground truth by hiring domain experts who manually detected spams given a few intuitive features [9, 10]. The authors observed some abnormal behavior regarding spam, and they classified the typical behavioral features of opinion spam and spammers into nine indicators.

While existing efforts discussed above present promising results, it is often easy for spammers to avoid content-based spamming detection by making superficial alterations to their reviews [10, 17]. Also, such pure content-based detection methods often need to develop different classifiers for each purpose and domain [12, 13]. By contrast, our approach detects spammers by analyzing user relationships and communities built through unusual interactions; which is much harder to fake than to reword their review content, as we shall describe later in this paper.

3 Review System

We focus on two types of user actions in a review system: reviewing items as *a reviewer* and commenting on reviews as *a commenter*. A *review* refers to one's opinion towards

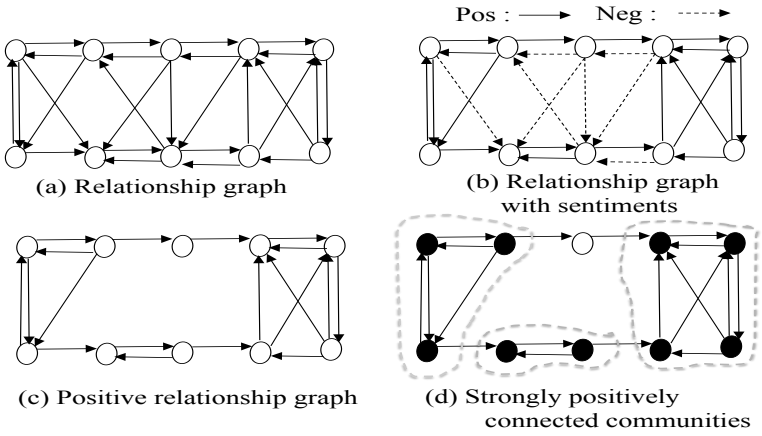

Fig. 1. The general idea of the proposed approach

an item and a *comment* refers to one's response to a review or other comments. Both reviews and comments may take a variety of forms including assigning scores, voting, and writing text. For instance, Amazon users can assign a star rating along with text reviews, post text comments, and vote on the helpfulness of a review; while Urbanspoon users can only vote on the helpfulness of a review but cannot post text comments.

Interactions are defined between two different users. More specifically, an interaction from a user u to a user v is formed if u made a comment on v's review or v's comment. Note that users may build threads of comments. In this paper, however, we will only consider the interactions made by commenting on a review for simplicity reasons. Also, we count multiple comments from the same user on the same review as a single interaction for the sake of fairness.

4 Discovering Opinion Spammer Groups

Our proposed approach aims to detect opinion spammer groups who artificially form communities through the coordinated positive interactions.

Generally speaking, our approach can be divided into four stages and Fig.1 depicts the general four stages through four sub-graphs, one sub-graph per stage. The four stages are: 1) building a general user relationship graph, Fig. 1(a); 2) annotating the general graph through sentiment analysis, Fig. 1(b); 3) pruning the general graph to a positive relationship graph, Fig. 1(c); and finally 4) identifying strongly positively connected communities within the positive relationship graph, Fig. 1(d). In the following, we will describe each stage in more details.

4.1 Stage 1: Building a General User Relationship Graph

We extend the definitions of a *user relationship* and a *community* proposed in our previous work [3], which we describe in this section.

We represent users and their interactions on a review system as a directed multi-graph $G = (U, E)$ in which U represents users (vertices) and E represents interactions (edges). Each edge $\overrightarrow{e_{uv}}$ is a 2-tuple (u, v) having direction from a commenter u to a reviewer v. A commenter has outgoing edges, and a reviewer has incoming edges in a graph. An *out-degree* of commenter u is the total number of edges from u to other users, which essentially reflects u's tendency as a commenter (i.e., how much u is willing to comment); an *in-degree* of reviewer v is the total number of edges from other users to v, which essentially reflects v's tendency as a reviewer (i.e., how popular v is to get comments). We further model those tendencies using incoming and outgoing probabilities defined as a reviewer's probability to get incoming edges and a commenter's probability to generate outgoing edges respectively.

Generally speaking, if we assume that there is no external relationship between commenter u and reviewer v, the typical interaction between u and v can be modeled as a random process. u simply stumbles upon v's review by chance when browsing the system. He does not know v and seek out v's review deliberately. In other words, without prior relationship from u to v, interactions from u to v should happen randomly depending on u's tendency as a commenter and v's tendency as a reviewer. Accordingly, we can represent all users interactions as a random graph in which edges (i.e., interactions) are randomly created following the incoming/outgoing probability of each user [3, 11]. Specifically, if the outgoing probability of u is p and the incoming probability of v is q in an original interaction graph G, $\overrightarrow{e_{uv}}$ is built with the probability $p * q$ in a random graph. Hence, we get a random graph $G_r = (U, E')$ in which the total number of all edges, and each user's degree distribution are the same as G. The random graph thereby preserves the same nature of each individual as a commenter or a reviewer, which is independent of any prior relationships between users. The only main difference between the two graphs is that: all edges are randomly generated in G_r and so the number of edges between each pair of users will be different from G.

Given the random graph model, we examine the real interaction patterns and see how much they are deviated from the random graph. We define users' relationship and its strength based upon the distance between users' original interaction graph and its corresponding random graph. Intuitively, the larger the difference between the real interaction pattern and the random model is, the more likely the relationships are to have been artificially orchestrated. We measure the distance by building confidence intervals based on the random graph. We denote that u has a relationship with v with τ strength, when the probability for edge $\overrightarrow{e_{uv}}$ to form in the real graph is outside of the given confidence interval τ. Then, the larger τ a relationship has, the farther the real interaction pattern is from the random graph and thus the higher strength the relationship has.

The concept of strength can be naturally extended to communities. Concretely, edge $\overrightarrow{e_{uv}}$ (in turn, user u and v) belongs to *τ% community*, if the strength of a relationship from u to v is τ. The larger τ is, the higher strength relationships in a community have and thus the higher strength the community has. Details of definitions can be found in our previous work [3].

For this work it is important to note that relationships belonging to higher strength of communities are excluded from lower ones. For instance, if a relationship is in 99.5% community, it is excluded from all lower strength of communities such as 98%.

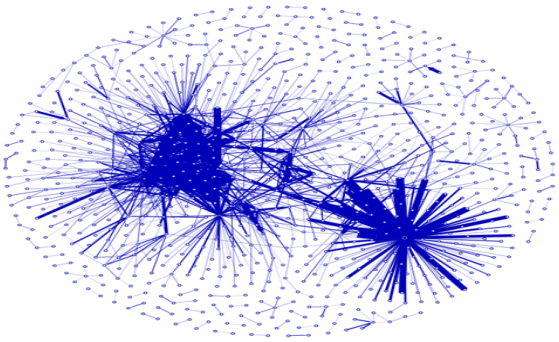

Fig. 2. An example of 99.5 % strength of user relationship graphs found in Amazon

Given the definitions above, we extract separate *user relationship graphs* for each τ community in which vertices are users and edges are their relationships defined by interactions, as illustrated in Fig. 1(a). Fig. 2 presents an example of 99.5% strength of user relationship graph found in Amazon.

4.2 Stage 2: Sentiment Analysis on User Relationships

Given τ strength of user relationships graphs, we further analyze the sentiment of relationships. To do so, we aggregate the sentiments of all comments between any pair of users from which we derive the sentiment of each relationship.

If comments are in the form of explicit votes, it is straightforward to obtain sentiment values. However, in many systems including Amazon and Yelp, it is often unknown who made the vote, while only aggregated information is publicly available. For example, we may know a certain review got 50 positive votes total, but we cannot know who made those votes. We thus focus specifically on the commenting text in order to define the sentiment of a relationship. We chose to employ a publicly available tool, AlchemyAPI [1], for this purpose. AlchemyAPI is known to present high accuracy on identification of sentiments in various applications including reviews and tweets [14, 16], which gives us good confidence in using the API.

AlchemyAPI takes text contents as input, identifies a sentiment of the text contents, and output sentiment score. The score ranges from -1 to 1, where positive/negative scores represent the strength of positivity/negativity, and 0 means neutral.

There are many possible ways to derive the sentiment of a relationship from the sentiment of each comment. In this paper we employ a straightforward approach where the sentiment of a relationship from commenter u to reviewer v is the average of the sentiments of all u's comments on v's reviews. Specifically, to decide whether a relationship between users u and v is positive or negative, we first analyze the sentiments of all comments between u and v, and aggregate them. We then build relationship graphs in which sentiments of all relationships are analyzed, as illustrated in Fig. 1(b). We consider the relationship is positive if the average sentiment score is bigger than 0.

4.3 Stage 3: Positive Relationship Graphs

Once sentiments are analyzed, we prune the user relationship graphs to build τ strength of positive relationship graphs by extracting only positive relationships (Fig. 1(c)).

4.4 Stage 4: Identify Spammer Candidates by Decomposing Positive Relationship Graphs

We identify spammer candidates by analyzing community structures in τ strength of positive relationships graphs. Note that we are interested in spammer groups who work together, not individual spammers. As mentioned before, to boost their opinions, each spammer needs to collect a significant amount of positive interactions from others, usually her colluders; as it is expected that non-spammers rarely post positive comments to spam in general, whereas groups of spammers post positive comments to each other so that most of them can obtain a dominant position (i.e., reviewers whose opinions are believed to be trustworthy) in a system. In other words, interaction from a non-spammer to a spammer are not likely to appear in positive relationship graphs; and spammers will have strong interconnections through positive interactions. This motivates us to extract strongly connected components from positive relationship graphs. In other words, we believe that deliberate positive interactions among spammers will lead to the formation of strongly connected communities in a positive relationship graph, as illustrated in Fig. 1(d). Accordingly, we cast the problem of detecting opinion spammers as the problem of finding strongly positively connected communities.

A *strongly connected component* $G' = (V, E')$ is a subgraph of given graph $G = (U, E)$ such that there is a directed path in each direction between every pair of vertices u,$v \in V \subset U$. In our context, we define a *strongly positively connected community* $G' = (V, E')$ as follows.

Definition 1. *G' is a* strongly positively connected community*, if G' is a subgraph of positive relationship graph G such that*

i) $\exists$ at least two vertices in G'

ii) G' is a strongly connected component of G

iii) G' is maximal, i.e., $\nexists$ strongly connected component $H \subset G$ containing G'.

We find all strongly positively connected communities in each τ strength of positive relationship graph and consider them as possible spammer candidates.

As noted before, non-spammers may form natural communities due to their similar interest on items. For example, in a random fashion, u has a few positive interactions with v through multiple items. This is likely to happen because v may have reviewed similar items, and u may look at those items to buy so that multiple interactions from u to v occur. And natural communities can arise from such random relationships. On the other hand, spammers construct artificial non-random communities. We thus argue that we can differentiate natural and spammer communities by measuring the level of randomness within the relationships. By our definition in Section 4.1, the strength of the relationships captures such a level of randomness, which we have in fact shown in [3]. We show that how the strength and spammicity are correlated in Section 5.

Category	#items	#reviews	#comments	#reviewers	#commenters
Books	116,044	620,131	533,816	70,784	164,209
Movie	48,212	646,675	201,814	273,088	72,548
Electronics	35,992	542,085	128,876	424,519	72,636
Tools	22,019	229,794	32,489	151,642	21,977
Across	222,267	2,038,685	896,995	901,812	295,118

Table 1. Dataset

5 Experimental Results and Analysis

In this section we will present the characteristics of the *discovered reviewers* who appear in the strongly positively connected communities identified by our approach. Table.1 summarizes the dataset collected from four popular categories of Amazon reviews: Books, Movies, Electronics, and Tools. In our experiment, we investigated the characteristics of the discovered reviewers in each category individually and across the four categories. This is because spammers may launch attacks not only in specific categories but also across categories. We will refer to the cross-category dataset as *Across*. Due to space limits, in the following we primarily report results for the Across dataset. We note that the same observations were found for the category-specific datasets.

We will compare three groups of reviewers: the *discovered* reviewers identified by our approach, the *top* reviewers recognized by Amazon, and the *total* reviewers which includes all reviewers appearing in our dataset. The top reviewers set contained 10,000 top ranked reviewers recognized by Amazon and to be on the list, a reviewer needs to demonstrate credible resources of high-quality reviews. Since Amazon is a well-known system, we assume that most top reviewers are trustworthy. Based on this assumption, we focus primarily on comparing our discovered reviewers with Amazon's top reviewers to show that although the discovered reviewers can appear to be as "helpful" as top reviewers, they are strong spammer candidates.

We begin by presenting statistical characteristics of discovered reviewers in Section 5.1. Then, we compare the behavior of three groups of reviewers in terms of verified purchase ratio (Section 5.2) and spam indicators introduced in [9, 10] (Section 5.3). Finally, in Section 5.4 we illustrate the effectiveness of our community-based detection by comparing it to the state-of-the art content-based approach.

In following sections we will describe each of the measures used for our study and will present the results in a series of plots where: the x-axis demonstrates the strengths of the recovered communities; the y-axis presents the strength of the appropriate measure; and we have one line per reviewer group.

5.1 User Statistics

Fig.3 shows the number of reviewers in each discovered community. In the Across dataset, no communities were found with strengths $10\% \sim 40\%$ and 70%. So they are not presented in the following graphics.

First and most importantly, we measured the average number of reviews of three groups of reviewers as shown in Fig. 4. Fig. 4 shows that both discovered and top

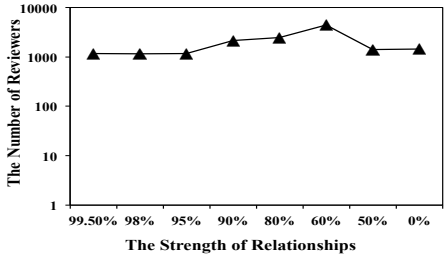

Fig. 3. The number of discovered reviewers

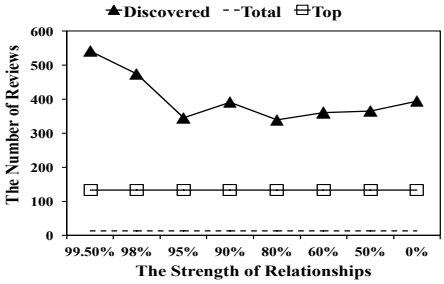

Fig. 4. The average number of reviews

reviewers have reviews more than 100 on average while the total reviewers have relatively less reviews on average, < 10. This result agrees with the observations of prior researchers who found that the majority of reviewers writes only a few reviews [18]. One important observation from Fig. 4 is that both the discovered and the top reviewers are *active reviewers* (i.e., who actively participate in discussion on items) and more importantly, our discovered reviewers are much more active on average than the top reviewers regardless of the strength of community they are in: > 300 reviews on average for discovered reviewers vs. 150 for the top reviewers. Additionally, the higher strength of communities (e.g., 99.5 % and 98 %) had more reviews on average, > 450, than those in the lower strength of communities (e.g., 0 %) $300 - 400$ range on average.

To determine whether the behavior differences reported in the following were due to the number of reviews a user submitted, we grouped the reviewers by the number of reviews they each submitted. We found that each group demonstrated similar patterns and behaviors as those for total population. Due to the space limit, we will not present the results here. However, this analysis showed the results presented in the following were not simply due to the reason that the discovered reviewers reviewed more.

As our goal is to find opinion spammesr who artificially boost their reviews, we first need to know whether their reviews are actually boosted in the system (i.e., whether their reviews are marked as helpful). In a common sense, reviews marked as helpful will have more influence on others. We thus calculated the positive vote ratio (PVR), ranging from 0 to 1, of the three groups of reviewers. We calculated PVR for each reviewer as the percentage of positive votes over the total number of votes each reviewer got. The higher PVR is, the more helpful their reviews appeared to be in general.

As shown in Fig.5, the PVRs of the discovered reviewers are relatively high and in fact, close to that of the top reviewers, nearly 80%. Both groups have much higher PVR than the total reviewers whose value is closer to 55%. This indicates that the opinions of discovered reviewers do indeed appear to be quite helpful in general, as much as that of the top reviewers. Additionally, PVRs of discovered reviewers vary across different strengths and 60% community has the lowest PVR ratio: close to 70%. In the following we show that although the PVR analysis indicates that reviews of discovered reviewers may have a similar level of influence on others as that of top reviewers, they are more likely to be spammers.

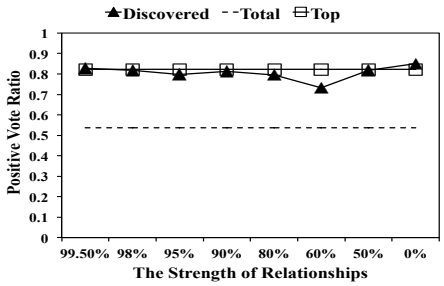

Fig. 5. Positive vote ratio (PVR)

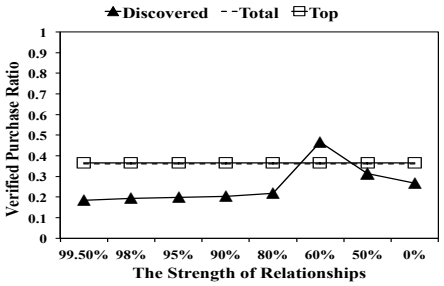

Fig. 6. Verified purchase ratio (VPR)

5.2 Verified Purchase Analysis

Amazon tags each review with Verified Purchase to indicate whether the reviewer made a purchase through Amazon. Although it is not the case that every non-spammer made a purchase through Amazon, reviewers who purchased the item are less likely to be spammers than those who submitted a review without doing so. We therefore defined the verified purchase ratio (VPR) as the percentage of verified reviews over the number of total reviews of each user and believe that VPR is good indicator for spammicity.

Fig. 6 shows the average VPRs of the three groups of reviewers. Interestingly, it also shows that there was no difference between the top and the total reviewers in terms of their VPRs. In other words, the top reviewers were no more likely to purchase the reviewed item than normal users. As we expected, our discovered reviewers have lower VPRs than the other two groups in general except for the 60% communities. In fact, the VPRs for the 80% $\sim$ 99.5% communities are substantially lower than those for the top and the total reviewers. For the reviewers in the 0% $\sim$ 60% communities we see that the stronger the community the higher the VPRs observed. However, as shown in Fig. 6, the trend is different for reviewers in the 80% $\sim$ 99.5% communities. In that case the strength of the community is negatively correlated with VPR. We believe that this occurs because the members of those communities are more likely to be spammers as we will show in the following sections.

5.3 Spammicity Analysis

In this subsection we will measure the spammicity (i.e., how likely users are to be spammers) of the discovered reviewers across the various community strengths. We used nine content-based spam indicators suggested by existing research to measure the level of spammicity of reviewers [9, 10]. Each value for the spam indicators ranges from 0 (non-spammers) to 1 (spammers).

Content similarity (CS): measures how similar the user's reviews are, as spammers often copy their own reviews across items. Following [9, 10], we measured the maximum of pairwise similarity of two reviews by each reviewer to capture the worst case. Fig.7 presents the average CSs of three groups of reviewers. Mukherjee *et al.* stated that the expected value of CS of spammers was 0.7 [9]. As shown in Fig.7, we observe that

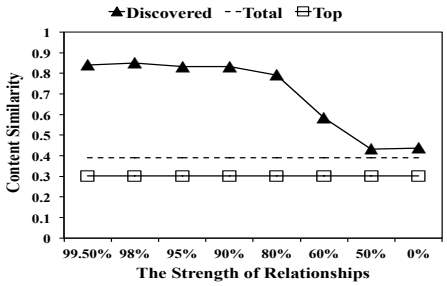

Fig. 7. Contents similarity (CS)

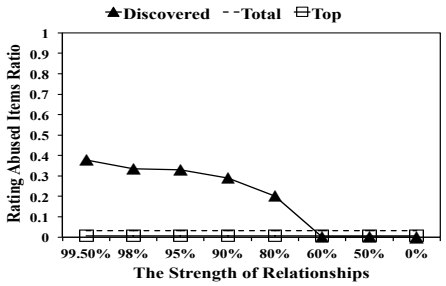

Fig. 8. Rating abused item ratio (RA)

the CSs of reviewers in 80% $\sim$ 99.5% communities are over 0.7. Note that there is a big drop between the 80% and 60% communities, and the CSs of 0% community is very close to the CSs of total reviewers. This result suggests that 80% $\sim$ 99.5% communities are more likely to be spammers with much higher CSs than others.

Rating abused item ratio (RA): checks whether a user posted multiple reviews with similar ratings on the same item, as non-spammers post multiple reviews usually when her opinion changes. Following [9, 10], we measured the similarity by computing the difference between the maximum and minimum ratings of each reviewer for an item; and we assumed a reviewer abused ratings, if she posted the similar ratings more than twice on the same item. We then measured how many items were involved in rating abuse for each user. Fig. 8 shows the average RAs of three groups of reviewers. In general, non-spammers are not likely to involve in rating abuse. Indeed, RAs of reviewers in 0% $\sim$ 60% communities and top reviewers are close to 0, whereas RAs of reviewers in 80% $\sim$ 99.5% communities range from 0.2 to 0.4 [3].

Maximum one day review ratio (MOR): measures how many reviews a user posted in one day compared with the maximum across all reviewers, as a massive amount of reviews in one day often looks suspicious. In the Across dataset, the maximum per day was 96, which we can undoubtedly say is a suspicious amount of reviews for a day. Fig. 9 shows the average MORs of three groups of reviewers. Mukherjee *et al.* stated the maximum number of reviews per day was 21 in their dataset, and the expected MOR of spammers was 0.28 and that of non-spammers was 0.11 (i.e., the expected number of reviews per day of spammers was 0.28 $\times$ 21 $\approx$ 5 and that of non-spammers was 0.11 $\times$ 21 $\approx$ 2) [9]. The maximum number of reviews per day was higher in our dataset than that used in [9] and this produced a correspondingly different MOR. However, we found that the maximum number of reviews per day ranged from 7 ($\approx$ 0.07 $\times$ 96) to 17 ($\approx$ 0.18 $\times$ 96) for reviewers in the 80% $\sim$ 99.5% communities, which is more than the expected number for spammers; whereas it was 3 ($\approx$ 0.03 $\times$ 96) for those in 0% $\sim$ 60% communities, which is similar to the expected number for non-spammers. It is interesting to see that the MOR of the top reviewers was also relatively high, compared

[3] In [9, 10], a measure called ***DUP (Duplicated Reviews)*** was also suggested, which focuses on multiple review content, not ratings. Our observations of the DUP were similar to our observations of RA. We therefore elected not to report the DUP due to space limitations.

to that of the total reviewers. One possible reason might be that Amazon invites some top reviewers to get advance access to not-yet-released items and to write reviews [2].

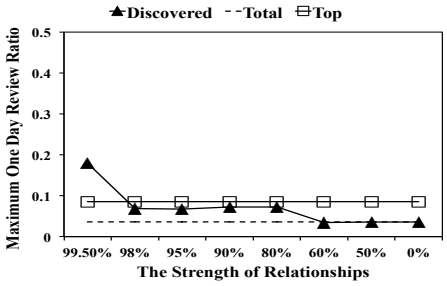

Fig. 9. Maximum one day review ratio (MOR)

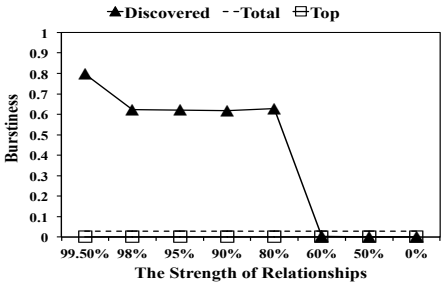

Fig. 10. Review Burstiness (BST)

Review burstiness (BST): measures the interval between a user's first and last reviews, as spammers often post reviews in a short period of time. Mukherjee *et al.* compared each reviewer's history with an estimated threshold of 28 days [9, 10]. The shorter the interval, the larger the BST. The BST was 0 if a reviewer has a history equal to or longer than 28 days. Fig. 10 shows the average BSTs of the three groups of reviewers. Note that top reviewers are expected to be valued customers who have a relatively long history with high-quality reviews. Indeed, top reviewers have the lowest BSTs (close to 0) as shown in Fig.10. By contrast, we observe that reviewers in the 80% and 99.5% communities have rather high BST scores. Recall that both the top reviewers and the reviewers in the 80% and 99.5% communities have high PVRs, but the BST score analysis suggests that the latter are likely to be spammers since do not have a long history but collect many positive comments in a short period of time to appear to be very "helpful".

First review ratio (FRR): measures how many of user's reviews are the first review for the target item, as spammers often post reviews early in order to maximize the impact of their reviews. Fig. 11 presents the average FRRs of three groups of reviewers. As shown in Fig.11, the top and the total reviewers have very close FRRs overall but for our discovered reviewers, we observe that FRR increases, as the strength of a community increases. Note that this result may simply reflect the fact that reviewers in the higher strength of communities are more active and thus are more likely to author the first review. However, the high FRRs for reviewers in 80% $\sim$ 99.5% communities still reflect their spammicity, when combined with other spam indicators [4].

Deviated rating ratio (DEV): checks the difference between a user's rating and the average rating of others for the same item, as spammers often try to inflict incorrect projections which deviate from the common consensus. We employed a threshold (of 0.63) estimated in [9, 10] to decide whether a rating is deviated, and measured the percentage

[4] In [9, 10], a measure called ***ETF (Early Time Frame)*** has also been suggested. The intuition behind this is the same as for the FRR, because if not for the first review, earlier reviews may have a bigger impact. Our observations of the ETF were similar to our observations of FRR.

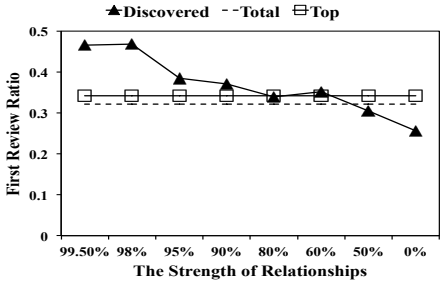

Fig. 11. First review ratio (FRR)

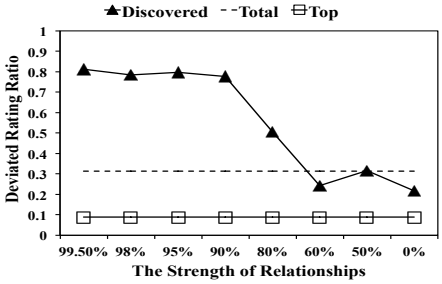

Fig. 12. Deviated rating ratio (DEV)

of a user's reviews deviated. Fig. 12 shows the average DEVs of the three groups of reviewers. Note that DEV of the top reviewers is the lowest. This suggests that their reviews are actually reliable or consistent with others' perceptions, whereas most reviews by reviewers in the 80% $\sim$ 99.5% communities deviate greatly from the common consensus. This deviance reaches as high as 0.8 deviation for the 99.5% community[5].

Summary In short, our findings from spammicity analysis can be summarized as follows. First, we find a clear distinction in terms of spammicity values between reviewers in the 80% $\sim$ 99.5% communities and those in the 0% $\sim$ 60% communities. Concretely, the behavior of the former groups tends to exhibit strong spamming behavior (high spammicity) although their positive vote ratio is high. The latter groups by contrast tend to be similar to the total and top reviewers (low spammicity). This result suggests that there exist reviewers whose reviews are maliciously endorsed to make more them influential. Indeed, prior researchers have argued that votes from users are not reliable and easy to abuse [7, 9].

Second, we see that the spammicity increases, as the strength increases for reviewers in the 80% $\sim$ 99.5% communities. In other words, reviewers in the higher strength communities (e.g., 99.5%) have a higher probability of being spammers; whereas reviewers in 0% $\sim$ 60% communities tend to have low spammicity in general, although the spammicity scores vary.

5.4 Analysis on Spammer Classification

In this section we show the correlation between the strength of each community and the probability of being spammers. Our goal is to suggest a way to incorporate distinctive characteristics of different strengths of communities for spammer detection with the analysis of false positive rate and true positive rate.

The most direct way to evaluate our approach is to compare our detection process to the state-of-the-art content-based scheme with ground-truth data. However, after several attempts, we were unable to obtain access to datasets with ground-truth labels used

[5] In [9, 10], a measure called ***EXT (Extreme rating ratio)*** was also suggested to determine whether a user's rating is extreme, as spammers' ratings tend to be extreme while that of non-spammers tend to be more moderate and item specific

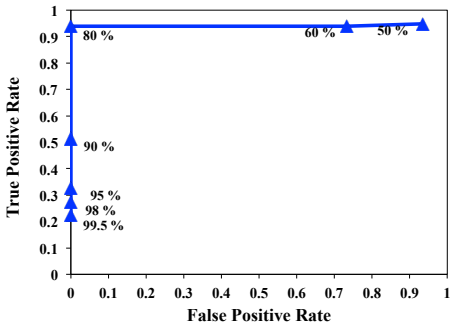

Fig. 13. ROC curve with different strengths as thresholds for classification

in previous research such as [9, 17]. Therefore we opted to compare our approach to the existing state of the art content-based classifier [9] on our dataset. Specifically, Mukherjee *et al.* suggested that when reviewers are ranked in descending order based on the sum of nine spammicity values, top and bottom 5% ranked reviewers can be classified as spammers and non-spammers respectively [9, 10]. We thus generated a "pseudo ground truth set" by applying Mukherjee *et al.*'s approach to our dataset, as it is the state-of-art classifier shown to have high accuracy over Amazon dataset with ground truth. We thereby show that although the proposed community-based approach does not look into details on review contents, it successfully identifies spammers without loss of accuracy.

Fig.13 shows the ROC curve by varying different strengths as thresholds to define spammers. The x-axis represents the false positive rate and the y-axis represents true positive rate. Each point represents true positive rate against false positive rate given τ strength as a threshold. We assume that reviewers in communities with strengths greater than or equal to τ are spammers; those in communities with strengths less than τ are non-spammers. For example, a point labelled as 90% represents that we assume reviewers in 90% $\sim$ 99.5% communities are spammers and those in 0% $\sim$ 80% communities are non-spammers. Note that in Fig.13, we present results regarding whether or not discovered reviewers in different strengths of communities are spammers. We thus do not plot when 0% is used as a threshold, as we could not get false or true negative results.

When 80% $\sim$ 99.5% are used as thresholds, there was no false positive as shown in Fig.13. This means that all reviewers in 80% $\sim$ 99.5% communities appeared in the top 5% ranked reviewers (i.e., spammers); which is expected, as their spammicity values were high as discussed in Section 5.3. Note that the larger threshold τ, the lower true positive rate. For example, when 99.5% is used as threshold, true positive rate is 0.2 due to many false negative results including reviewers in 80% $\sim$ 98% communities. On the other hand, when 50% or 60% was used as thresholds, the false positive rate dramatically increased (over 0.7), meaning that 0% $\sim$ 60% communities are likely to be non-spammers. As more non-spammers (i.e., reviewers in 0% $\sim$ 60% communities) are classified as spammers with 50% or 60% as thresholds, the number of false positive results increased. In such a case, the number of false negative results would be small, resulting in higher true positive rate.

Note that we get the best result (i.e., 0 % false positive rate and high (close to 1) true positive rate), when 80% is used as a threshold; and the classifying results get worse with thresholds lower than 80%. This implies a clear distinction between reviewers in $80\% \sim 99.5\%$ communities and those in $0\% \sim 60\%$ communities. This lends support to our claim that strengths of communities can be used to distinguish spam communities from non-spam ones.

Summary In short, our findings from ROC analysis indicate that while strongly positively connected communities may be naturally constructed with different strengths, communities with a strength higher than 80% are strong spammer candidates. We note that it is hard to evade our community-based scheme as spammers essentially need to build such high strength of communities to make their opinions influential; whereas spammers can easily fake their content features (e.g., reword their contents to lower content similarity value) to evade detection by content-based classifiers. It is also important to note that discovered communities not only include reviewers but also commenters who may not write any spam reviews. Existing pure content-based approaches will not be able to discover such *supporting commenters*, though they are also suspicious and indirectly contribute to opinion spams. In other words, our approach can discover both spam reviewers and suspicious commenters, which is a great advantage over pure content-based approaches.

6 Conclusion

In this paper we proposed a novel approach to find opinion spammer groups by analyzing community structures built through abnormally non-random positive interactions based on the intuition that spammers need to form artificial communities to make their opinions influential. We thereby exposed two types of spammers: spam reviewers who post spam reviews and supporting commenters who extensively endorse those reviews. Through extensive experimental analysis, we demonstrated the effectiveness of our community-based approach in terms of accuracy and reliability. We showed that our approach can successfully identify without relying on review contents, while achieving the same level of accuracy as the state-of-art pure content-based classifier.

Some challenges still must be surmounted. First, the proposed approach has focused mainly on spammer groups so it cannot find individual non-group spammers. We may combine our approach with content-based classifiers (e.g., [9, 18]) to detect such non-group spammers. Second, while we have discussed the effectiveness of our approach in terms of detection accuracy, it would also be useful to develop a model to measure the effect of various spamming strategies (e.g., manipulate contents and build artificial communities). We thereby plan to investigate the robustness of our approach (i.e., to what degree attackers can manipulate their behavior to avoid detection).

Acknowledgement

This work is supported in part by the National Science Foundation under the awards CNS-0747247, CCF-0914946 and CNS-1314229, and by an NSA Science of Security

Lablet grant at North Carolina State University. We would also like to thank the anonymous reviewers for their valuable feedback.

References

[1] AlchemyAPI: Http://www.alchemyapi.com/
[2] AmazonVine: Http://www.amazon.com/gp/vine/help
[3] Choo, E., Yu, T., Chi, M., Sun, Y.: Revealing and incorporating implicit communities to improve recommender systems. In: Proc. of the 15th ACM Conf. on Economics and computation. pp. 489–506. ACM (2014)
[4] Feng, S., Xing, L., Gogar, A., Choi, Y.: Distributional footprints of deceptive product reviews. In: ICWSM (2012)
[5] Jindal, N., Liu, B.: Opinion spam and analysis. In: Proc. of the WSDM. pp. 219–230. ACM (2008)
[6] Jindal, N., Liu, B., Lim, E.P.: Finding unusual review patterns using unexpected rules. In: Proc. of the 19th CIKM. pp. 1549–1552. ACM (2010)
[7] Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., Lauw, H.W.: Detecting product review spammers using rating behaviors. In: Proc. of the 19th CIKM. pp. 939–948. ACM (2010)
[8] Lu, Y., Zhang, L., Xiao, Y., Li, Y.: Simultaneously detecting fake reviews and review spammers using factor graph model. In: Proc. of the 5th WebSci. pp. 225–233. ACM (2013)
[9] Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R.: Spotting opinion spammers using behavioral footprints. In: Proc. of the 19th ACM KDD. pp. 632–640. ACM (2013)
[10] Mukherjee, A., Liu, B., Glance, N.: Spotting fake reviewer groups in consumer reviews. In: Proc. of the 21st WWW. pp. 191–200. ACM (2012)
[11] Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E 64, 026118 (Jul 2001)
[12] Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: HLT-NAACL. pp. 497–501 (2013)
[13] Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557 (2011)
[14] Quercia, D., Askham, H., Crowcroft, J.: Tweetlda: supervised topic classification and link prediction in twitter. In: Proc. of the 3rd WebSci. pp. 247–250. ACM
[15] Sheibani, A.A.: Opinion mining and opinion spam: A literature review focusing on product reviews. In: 6th IST. pp. 1109–1113. IEEE (2012)
[16] Singh, V., Piryani, R., Uddin, A., Waila, P.: Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In: Proc. of iMac4s. pp. 712–717. IEEE (2013)
[17] Wang, G., Xie, S., Liu, B., Yu, P.S.: Review graph based online store review spammer detection. In: 11th ICDM. pp. 1242–1247. IEEE (2011)
[18] Xie, S., Wang, G., Lin, S., Yu, P.S.: Review spam detection via temporal pattern discovery. In: Proc. of the 18th SIGKDD. pp. 823–831. ACM (2012)
[19] Yoo, K.H., Gretzel, U.: Comparison of deceptive and truthful travel reviews. Information and communication technologies in tourism 2009 pp. 37–47 (2009)