공학석사학위논문

의견 스패머에 대응하는 사용자 행동 분석: 동조 및 자정 효과

Analysis of User Behavior against Opinion Spammer: Following and Correcting Effect

2018년 8월

서울대학교 대학원
컴퓨터공학부

박 지 현

의견 스패머에 대응하는 사용자 행동 분석: 동조 및 자정 효과

Analysis of User Behavior against Opinion Spammer: Following and Correcting Effect

지도교수 김 종 권

이 논문을 공학석사학위논문으로 제출함

2018년 4월

서울대학교 대학원

컴퓨터공학부

박 지 현

박지현의 석사학위논문을 인준함

2018년 6월

위 원 장 전 화 숙 (인)

부 위 원 장 김 종 권 (인)

위 원 이 상 구 (인)

ABSTRACT

Analysis of User Behavior against Opinion Spammer: Following and Correcting Effect

Jihyun Park
Dept. of Computer Science and Engineering
The Graduate School
Seoul National University

Opinion Spam is still a widespread problem in online review platforms. Opinion spam is hard to detect, because of spammers' sophisticated strategy to avoid detection system. In contrast with other kind of spams, context is not a powerful feature to detect opinion spam. This is the most challenging point of detecting opinion spam. In this thesis, we analyzed opinion spams' effect from the perspective of truthful users' reactions. We found out a timing of spammers' attack and also showed an activity of users is increased after the attack. Normally users agreed with previous reviewers' opinion and we observed the phenomenon became more evident when

spammers attacked a product. And we found out that there are both following action and correcting action of truthful users who are affected by spammers. After the attack, some of truthful users are hooked to spammers and follow them, whereas, some others try to remedy contaminated online society. We used Yelp dataset to analyze temporal dynamics around spammers and revealed these significant signals with empirical and statistical probability. This is the first research analyzed truthful users' action responding to opinion spammers. We also detected spammers' attack period using our new observations. As a result, we identified spammers' attack strategy, effect, discovered truthful users' action patterns (following and correcting effect) responding to spammers, and effectively detected spammers' attack period.

Keywords: Opinion Spam, Review Spam, User Behavior Analysis, Opinion Spam Detection, Social Influence, Social Network, Correcting Effect, Following Effect

Student Number: 2016-21206

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter I

Introduction

1.1 Background

As the online market grows, the influence of reviews is also growing. In addition, as the size of the online review platforms grows, reviews play an important role not only in making a decision what to buy in e-commerce, but also what to eat, where to stay, and where to go. Since the scope of reviews' effect is broadening, and people are dependent on reviews before visiting, business owners should always be mindful of their reputation on the review platforms.

Reviews are a measure of a stores' popularity. The greater the number of reviews, and the higher the rating, the greater chance of exposure to review site users. As the success of a business becomes more and more dependent on the users' reviews, the owners who run the store deeply understand the importance of reviews. Thus, while owners often use a healthy marketing method to drive reviews, some owners also pay for fake reviews.

A fake review is written to promote or defame a product or store, not a frank review of the user after actual purchase. Specifically, it refers to the act of leaving a review that satisfies the conditions requested by the clients (i.e. store owners). These fake reviews are called *opinion spam* [1]. Crowdsourcing websites raise owners (client) who need a promotion and pay for crowd workers who wrote fake

reviews to the client. We will call *promoting spam* to increase a reputation, and *defaming spam* to hurt a reputation. *Promoting* spam is more general case compare to *defaming* spam, we focused on *promoting* spam.

Opinion Spam is one of the most complex spam to detect it. In the case of web spam or spam mail, since it is a machine-generated spam, text information could be used sufficiently for a detection [2]. However, the case of opinion spam, high-quality reviews from real human workers, is hard to detect only with context information. For this reason, although at the beginning of studies researchers used linguistic features [3,4], using spammers' behavioral feature become one of the most common approach to build a detection model [4,6,7]. Moreover, they found temporal patterns [8,9], and construct network [10,11] or tensor [12,13] to detect spammers or spammer group [14].

1.2 Goal and Contribution

Until now, researches related to opinion spam only concentrated on spammer itself. Yet, no one focused on the temporal dynamics around (before and after) spammers' attack even though there may be have indicative signals of the attack. What happens before the spammers act? What about the situation after the attack? Are these spammers really effective? How will normal users react? We analyzed these questions aspect of a temporal behavior of normal users who react to spammers. We use YelpNYC dataset and select targeted products and pure products to compare users' behavior between them.

Through this study, we first catch the behaviors of normal users

who showed following action to spam activities, and users who rather acted against spammers' opinion and self-corrected a contaminated public opinion. If normal users follow the spammers' fake opinion, the owner will be worth hiring spammers, but rather try to curb spammers' activities, there will be opposite effect which owner didn't expect. We show both of these behaviors statistically. And finally, we detect spammers' attack period using these observation and show high performed results.

The main contributions of this thesis are as follows:

- We generally analyze opinion spammers' strategy and effect.
- We observe herding behavior between same opinion users.
- We generally investigate normal users' temporal behaviors responding to spammers using novel measurement.
- We observe *following and correcting* effect after spammer's attack.
- We detect spam attack period using these observations.

1.3 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 describes the related work of opinion spam detection and social influence. And Chapter 3 gives a description of dataset we used. Main analysis and observations are described in Chapter 4. We show herding behavior of users (Chapter 4.1) in the first part of this chapter. And analyzed strategy and effect of spammers (Chapter 4.2). And then explain two

type of truthful(non-spam) user's behavior responding to spammer's action (Chapter 4.3). After the analysis, we detected spam attack period in Chapter 5. Finally, Chapter 6 concludes this thesis.

Chapter Ⅱ

Related Work

2.1 Opinion Spam Detection

Papers related to opinion spam usually detect opinion spam or observe new features. Opinion spam detection was introduced in [1]. In the early stage, textual elements [3,4] were used, but the reviewer's behavioral features [5,6,7] are typically used instead of linguistic features. Textual information is almost useless since a human exquisitely writes opinion spam. Rating deviation, rating distribution, entropy, review's length, and singleton review or not [9,11,14,15] are generally used as distinct behavioral features of spammers.

[11,13,16,17] construct a network, and by combining behavioral feature, they increase the performance. Using network features for detection have a benefit of utilizing neighbor nodes' information. There is a method to detect spammer groups on the assumption that spammers conduct a campaign and collaborate with each other. Recent papers use time-series data actively to catch the group. Some detect burst interval [8,9,14,15,18] and some others detect dense-block [12,13,14,19] to detect spammer groups.

[8] observes spammers' strategies based on temporal dynamics of

truthful ratings. [9] detects products' peak point from Kernel Density Estimation, and conducts a network of users who had appeared in burst interval. And they finally detect spammers using Markov Random Field. [18] analyzes that spammers' inter-arrival time has bimodal distribution and detects co-bursting spammer group.

Tensor-based method M-Zoom[12] and D-Cube[13] detect spammer group who shows a lockstep behavior. They construct tensor using time information and IP. HoloScope[14] also identifies spammer group using graph topology. They find a group which maximizes the difference of features between a standard group. They detect burst interval and use rating deviation feature. However, they use spammers' expected trait without any observations.

A lot of studies detected spammers based on assumptions of spammers' behavior itself. Assumptions commonly used is that a spammers' group activity occurs burstiness. As previous papers have shown that spammers have several patterns in their movement, truthful users can also have patterns. Through this thesis, we indicate that burstiness is not only occurred by spammers but also truthful users who followed by them. This observation can amplify the signal of attack.

As opinion spam filtering system becomes more sophisticated, spammers also develop to complicated strategies and try to act normal to avoid detection. However, the only thing that they can not invent is resulting truthful reviews. Even if the spammers camouflage well, following reviews can give indicial evidence of spammers' activity.

2.2 Social Influence

In this thesis, we try to analyze a relationship between spammers and honest users which is not discovered yet. We show the correlation from a social influence point of view. Social influence is that others' action or opinion influence one. Herding behavior or information cascading explain social influence.

Herding behavior [20] is that users follow others' previous ratings of the same product. And information cascade [21] means not only one follows public opinion (like herd behavior) but follows what others do even if one has his private opinion.

[22] studies a social influence of prior ratings by crowds and friends based on information cascades theory. And they have shown social networking (like friends' rating) reduces herding behavior by the crowd. In [23], they investigate an impact of social influence on the success of a poll. They show that early respondents have a larger influence and respondents are influenced the most by others who posted just before them. [24] also discovers average ratings of the same product is different on different websites due to the herding effect.

[25] observes the assimilation and contrast effects in users' rating behavior caused by historical ratings. Assimilation effect occurs if a products' quality and history ratings are similar, and if they have a deviation, contrast effect has occurred. They model a real rating system and predict following scores.

Unlike previous papers, we focus on the phenomenon caused by

spammers rather than the aspects between ordinary users. Previous studies have also shown that historical ratings and opinion affect the following users, but no one yet investigated whether spammers' activities lead to the same behavior for the following truthful users. Do spammers change the action of other users? Does the rating go as spammers induce? Or are others correcting it? In this thesis, we show not the behavior of spammers themselves, but a social influence resulting from them.

Chapter Ⅲ

Dataset

Opinion spam is difficult to detect only with public information unlike other kinds of spams. So, accurate detection could be achieved from internal information which is accessible inside of the website [26, 27]. Several review platforms (e.g., Amazon, Yelp, TripAdvisor, DianPing, and so on) developed their opinion spam filtering system and provided a quite exact result under vast information. However, they never open to the public how did they find spammers because they have to protect their system from spammers who can escape their algorithm.

As considered in [27], Yelp exploits behavioral features extracted from its' internal data (e.g., IP addresses, click behaviors, network logs, geolocations, and so on) for detection and its' result is reliable. Like previous studies [8, 11, 28], we can use reviews filtered by Yelp as spam (near ground truth).

We use YelpNYC dataset, restaurant data in New York City, used in [11]. The data include {Product ID, Reviewer ID, Rating, Date, Content, Label} information. We selected 124 products (targeted product) which spam rate is over 15% of 923 products. Among the regular (truthful) reviews, we refer 1-3 rating reviews as defaming real ("*dReal*") and 4-5 rating reviews as promoting real ("*pReal*"). The reason why a rating score three can be treated as a low is that

the average of each of the 124 targeted products is larger than 3. And we refer 4-5 rating reviews of Yelps' filtered reviews as promoting spam ("*pSpam*"). 1-3 rating reviews among filtered reviews ("*dSpam*") take less than 5% of all reviews, so we excluded in this thesis. Therefore, we categorized review with 3 type("*pReal*", "*dReal*", "*pSpam*"). Table 3.1 summarizes the dataset.

Table 3.1: Statistics of dataset used in this work

	Targeted Product	Pure Product
# Products	124	94
# *pReal* reviews (%)	17,351 (65.15%)	16,789 (73.43%)
# *dReal* reviews (%)	4,101 (15.40%)	5,020 (21.95%)
# *pSpam* reviews (%)	4,106 (15.42%)	758 (3.32%)
# *dSpam* reviews (%)	1,075 (4.03%)	298 (1.3%)
# Reviews	26,633	22,865
# Reviewers	22,917	17,023

Chapter Ⅳ

Empirical Measurement and Observations

We analyze behaviors of truthful users followed by spammers using empirical measurements and discover six novel observations. Patterns we investigated are described as follows:

- In online review platforms, users of the same opinion act together. And this phenomenon occurs more strongly by spammers.
- A store owner employs spammers when the reputation fells down.
- Spammers' activity activates movement of users.
- After spammers' activity, *following* effect occurs as the owner expected.
- After spammers' activity, *correcting* effect occurs in contrast with what the owners' thought.
- Correction review acts when the same opinions exist in front of it.

We observe these phenomena generally depending on a statistical baseline. To do this, we propose and use a new measurement method. Above observations show, owing to the spammers, owners can have a more significant effect than they expected or the opposite effect due to the public trying to correct a wrong agitated opinion.

From now on, we will explain how we found each pattern and what method we used. Section 4.1 shows common behavior on all type of reviews, and 4.2 shows strategy and effect of spammers. Section 4.3 finally indicates truthful users' induced activity by spammers. And notations used in this thesis are defined in Table 4.1.

Table 4.1: Notations

Symbol	Definition
$A^t = \{(a_1^t, i_1^t, p_1^t), (a_2^t, i_2^t, p_2^t), \cdots, (a_{N^t}^t, i_{N^t}^t, p_{N^t}^t)\}$	Triple set of t type reviews
$A_p^t = \{(a_1^t, i_1^t, p), (a_2^t, i_2^t, p), \cdots, (a_{N_p^t}^t, i_{N_p^t}^t, p)\}$	Triple set of t type reviews for product p
$t \in \{pReal, dReal, pSpam\}$	Type of review
a_n^t	nth anchor review in set A^t
p_n^t	A target product of review a_n^t
P	Set of products
i_n^t	a_n^t's position in a product p_n^t's chronologically ordered reviews
N^t	# of t type reviews
N_p^t , N_p	# of t type reviews for product p, # of all reviews for product p
$T_p(k)$	kth positions' review type for product p
$R_p = \{(r_1, d_1), (r_2, d_2), ..., (r_{N_p}, d_{N_p})\}$	Pair set of time-sorted reviews for product p
$R_p(i) = \{(r_i, d_i), (r_{i+1}, d_{i+1}), ..., (r_{i+W-1}, d_{i+W-1})\}$	W sized subset of R_p which index starts at i
r_i	ith review in R_p
d_i	Written date of review r_i
W, W_f, W_l	Window Size, (former, latter)-Window Size
$D^{(t,t',k)}$	Difference of Appearance Prob.
$C_p^t(i)$	# of t type reviews on product ps' ith window
$G_p(i)$	Time Gap(interval) on product ps' ith window

4.1 Herding Behavior between the Same Opinion Users

An opinion of previous users is an essential factor for following users in online society. Even if there is another private opinion, they support public opinion especially right before of them [23]. We wanted to see if this property also presents on Yelp. Observing the herding behavior can be done by confirming how many users of the same opinions (same type of reviews) are working together.

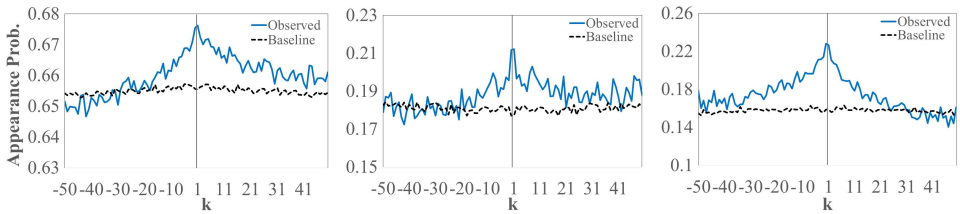
We proposed **appearance probability** which can see statistically whether grouping behavior is happened or not. Appearance probability is a value which represents how much of the same review type is appeared in k point (index) far from a specific review.

$$1_n^{(t,t')} = \begin{cases} 1, \text{if } T_{p_n^t}(i_n^t + k) = t' \\ 0, else \end{cases}$$

From a t type anchor review a_n^t, the indicator function $1_n^{(t,t')}$ is one if the review, k points far from the anchor, is t' type review. Accordingly, appearance probability can be calculated as follows:

$$\Pr(T(i^t + k) = t') = \frac{\sum_{n=1}^{N^t} 1_n^{(t,t')}}{N^t}$$

From all the t type anchor reviews (A^t), it is an empirical probability of appearance of t' type review which is k point away from anchor reviews. Appearance probability can be used to check whether users of the same opinion are clustered at similar positions. For comparison, we set up a baseline shuffling all the reviews' order of each product. (To make it general, we generated ten shuffled



(a) $t = t' = pReal$ (b) $t = t' = dReal$ (c) $t = t' = pSpam$

Figure 4.1: Appearance probability of the same type reviews. Gray vertical line is zero point. $k \in [-50, 50]$

datasets and averaged it.)

Figure 4.1 shows appearance probability for each type of reviews. We set type t' equal to t, which represent the probability of appearing the same type of reviews at k points far away. Gray vertical line in the middle is the point when k is 0. In other words, the left side of the line shows the probability of the same type of reviews before the anchor point, and the right side shows the probability after the anchor point. As can see in Figure 4.1, the likelihood of baselines is stable, whereas, in the observed probability, the smaller the $|k|$ is, the higher the probability. It means that the likelihood of appearing the same type of reviews around the anchor points is higher than at a distance.

We confirmed that truthful users appear in a similar position with the same opinion users. It can be interpreted as herding or information cascading behavior. The fact that spammers work together has a different meaning for truthful users. It is an evidence of a campaign which spammers attack together. Additionally, in Section 4.3.2, we captured that these group activity between the same opinion users become dominant due to spammers.

4.2 Strategy and Effect of Spammer

4.2.1 Attack after Low Score

Store owners are sensitive to their reputation since it directly affects the profits. They will be afraid of getting a low score from the reviewers. So owners track their review score continuously and will employ spammers if the score fell. It has already been shown that spammers work when a target store is badly evaluated by truthful reviewers [8]. However, [8] shows a normalized value for a specific product, so it is difficult to see exactly how significant changes occurred in general. Then, we discover that lousy evaluation activates spammers' activity using a difference of appearance probability.

A **difference of appearance probability** is proposed to show relative changes between observed appearance probability and statistical baseline. A statistical baseline for each product is the rate of each type of review in the product. We compare it to the observed value and show how much more appeared at each position.

The statistical probability of appearance of t' type review at any point i in a product p is calculated as:

$$\widehat{\Pr}_p(T(i)=t') = \frac{|A_p^{t'}|}{N_p}$$

The empirical probability of product ps' t' type review appearance at k index far from t type anchor review can be calculated as:

$$\Pr_p(T(i^t+k)=t') = \frac{\sum_{n=1}^{N_p^t} 1_n^{(t,t')}}{N_p^t}$$

We will refer them as $\widehat{\Pr}_p^{(t')}$, $\Pr_p^{(t,t',k)}$ each for simple.

And then, we convert the value as a relative one, a difference of appearance probability. It is computed as:

$$D_p(T(i^t+k)=t') = \frac{\Pr_p^{(t,t',k)} - \widehat{\Pr}_p^{(t')}}{\widehat{\Pr}_p^{(t')}} \times 100$$

It means that, in product p's time series data, from the t type anchor reviews, t' type review has occurred $D_p^{(t,t',k)}$ percent more than expected at k points away from each anchor.

To show it generally, we calculate the value for all products and average it.

$$D^{(t,t',k)} = \frac{\sum_{p}^{|P|} D_p^{(t,t',k)}}{|P|}$$

We saw the difference in the appearance of *pReal* type $D^{(pSpam,pReal,k)}$ and *dReal* type $D^{(pSpam,dReal,k)}$ with *pSpam* type review as an anchor point, respectively in Figure 4.2. We can treat a black vertical line as the point when spam has occurred. In the case of *pReal*, they acted less than expected right after/before spam appears. However, in the case of *dReal*, at the position right before spam appears ($-7 \le k \le -1$), they occasionally looked more than expected, contrasted to the value right after a *pSpam* appeared. With this figure, we discovered that spammers attacked more when low ratings

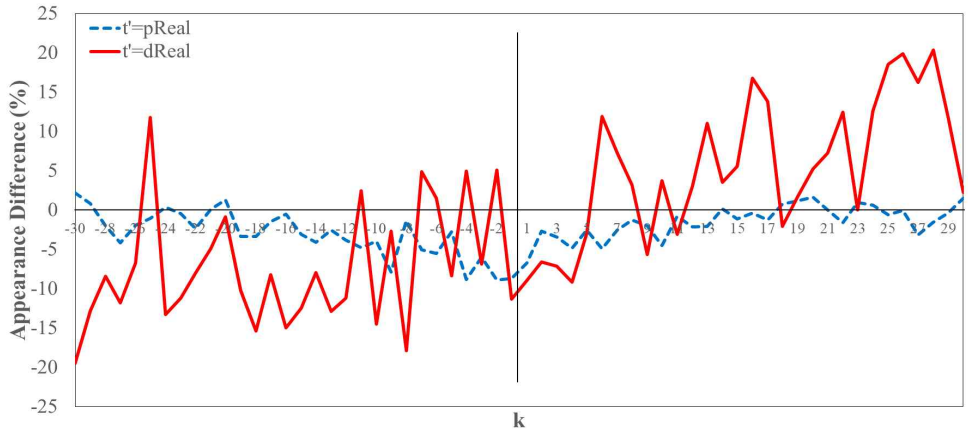


Figure 4.2: Difference of appearance probability. Anchor reviews are a *pSpam* type ($t = pSpam$). $k \in [-30, 30]$

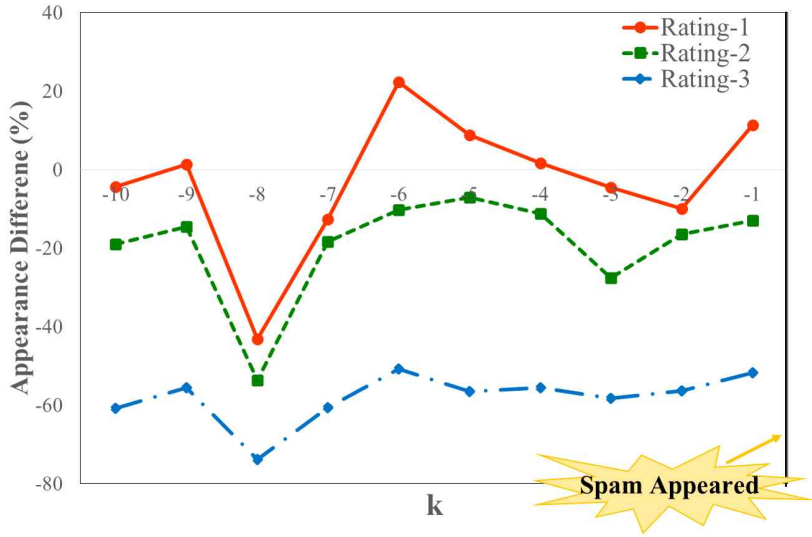


Figure 4.3: Difference of *dReal* appearance probability before spam appears. $k \in [-10, -1]$.

appeared than high scores.

To check more in detail, we divided *dReal* type in each rating (1, 2,

3). Figure 4.3 shows each ratings' appearance right before the spam appears. As we expected, the lower rating showed higher appearance probability. 1 rating score has a more significant impact on spammers than 2 or 3; thus spammers act right away to recover the reputation.

4.2.2 Activeness after Attack

We find out how will the activeness of users change after the attack. Figure 4.4 shows a time series of reviews' count of a product. Each value is calculated for two weeks. As can see, there was a steep rise in the number of reviews right after spams appeared (2015/05/23). Not only for the spammers but the truthful users.

To investigate it generally, we saw the period after the spam attacks strongly. Figure 4.6(a) shows a period get short as the rate of spams gets high. We calculate the rate of spams in 10 reviews at former-window($W_f = 10$), and the **period** is days it took to get 20 reviews at latter-window($W_l = 20$). We slide the windows for all products and average all calculated period for each observed spam rate. The result means that as spammers act more together, following reviews are written in shorter day gap.

We also observed an increase in activity depending on an analytical baseline (Figure 4.6(b)). We classified the activity of spammers in former-window and the period of latter-window according to the criteria. Since each product has a different rate of spammers and truthful users, we should set different activeness baseline for each product.

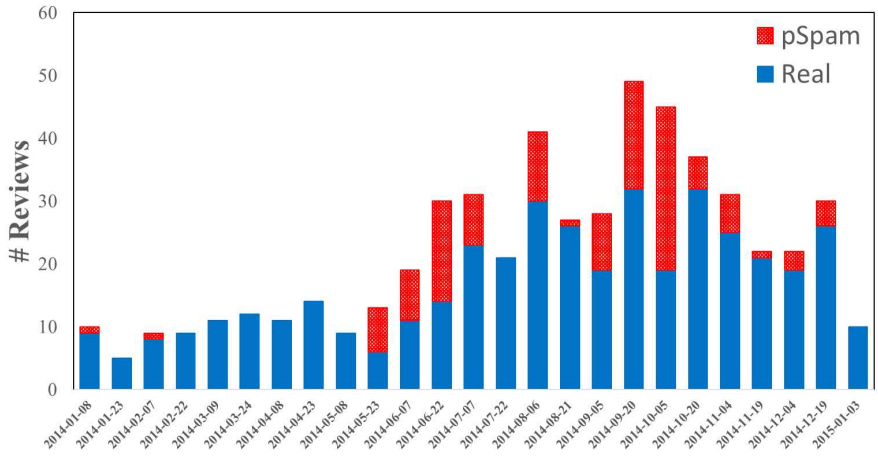


Figure 4.4: The number of reviews for two weeks

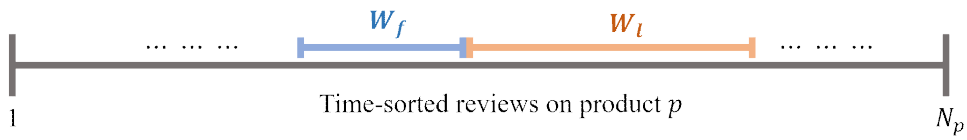


Figure 4.5: Definition of former-window and latter-window

First, we count the number of spam in former-window and compare to the expected amount of spam in the window. It can be calculated as follows:

$$C_p^S(i) = |\{r|r \in R_p(i), r : pSpamtype\}|$$

$$\hat{C}_p^S = \frac{|\{r|r \in R_p, r : pSpamtype\}|}{|R_p|} \times W$$

$C_p^S(i)$ refers to the observed number of *pSpam*("*S*") reviews in ith former- window of product p and $\hat{C}_p^S$ refers to the expected number of spam in a window.

And then, we set up the criteria whether a latter-window is "*active*"

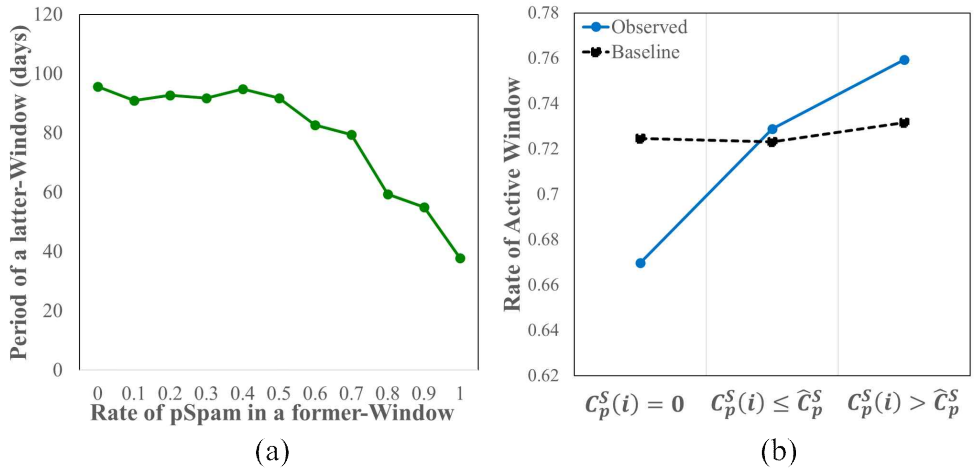


Figure 4.6: (a) Change of the period (b) Rate of active window, $W_f = 10$, $W_l = 20$

or not. ith window is *active* when the observed period($G_p(i)$) of a window is shorter than the expected period ($\hat{G}_p$).

$$G_p(i) = d_{i+W-1} - d_i$$

$$\hat{G}_p = \frac{(d_e - d_0)}{|R_p|} \times W$$

Overall, for ith former-window $R_p(i)$ in a product p, calculate $C_p^S(i)$ and compare to $\hat{C}_p^S$, and then calculate the following latter-window $R_p(i+W_f)$s' day gap $G_p(i+W_f)$ and check whether it is *active* or not.

Figure 4.6(b) shows how many *active* latter-window appeared according to a spams' concentration in former-window. Black dashed line is baseline which is a value of shuffled datasets (iterated ten times for generality). As can see, there was no difference in the baseline, whereas, real-world dataset showed a linear increase as a concentration of spam gets higher. Figure 4.6(a) and (b) shows the

effect of spammers attack in general; after the spammers' strong attacks, users write reviews more *active*.

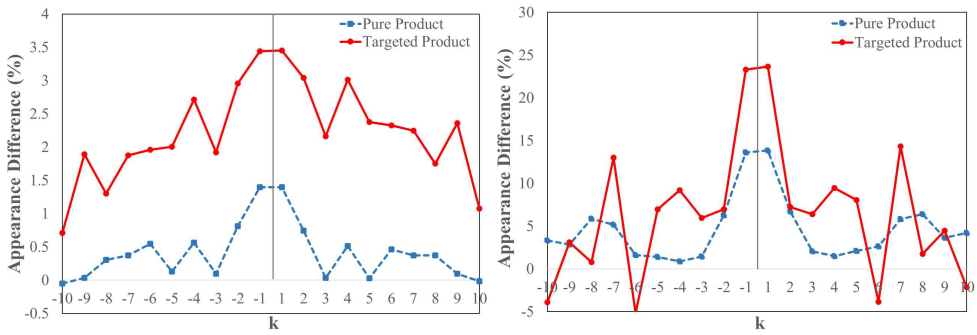
4.3 Truthful users' Behavior responding to Spammer

The main contribution of this thesis is that we observed a behavior of truthful users in response to spammers' attack. We found that as the spammers became active, users of the same opinions (*pReal-pReal*, *dReal-dReal*) gathered together. And also, two patterns emerged among truthful users; both *following* effect which sympathizes with spammers and *correcting* effect that rose up in spammers' opinion.

4.3.1 Strong Group Action of Truthful Users

In the previous chapter 4.1, we observed herding behavior on the same opinions of users (i.e., *pReal*, *dReal*, *pSpam*). And now, we find out what is the difference of the behavior on targeted products and non-targeted products(pure products). That is, the question is how the truthful users behave when spammers promote the product by a group. To compare the behavior, we select pure products which have less than 4.7% of spam. We selected 94 products from YelpNYC to make the number of truthful users comparable with targeted products (Table 3.1).

We used the difference of appearance probability to see how the



(a) $t = t' = pReal$ (b) $t = t' = dReal$

Figure 4.7: Difference of appearance probability for targeted / pure products.

same opinion works together on each dataset. In Figure 4.7(a), the anchor points are *pReal* type reviews and saw the appearance of *pReal* type reviews from ten points before to ten points after the anchor point ($t = t' = pReal$, $-10 \leq k \leq 10$). An aspect of the graph was similar, a high difference in low $|k|$, which means herding effect occurs both in targeted and pure products. However, since the value of targeted products is always more prominent than that of pure products, *pReal* reviews work more collectively in the targeted products. Same in *dReal* type reviews (Figure 4.7(b)). In the range of $-5 \leq k \leq 5$, the value was also high in small $|k|$ in both, even though, *dReal* ratings followed each other more in targeted products.

This result can be interpreted as the reaction to spam. Because of spammers, who intentionally manipulated the public, the users who are instigated by them (*following* effect) and those who kick against them (*correcting* effect), are both actively grouped. Based on the above results, we have verified in the next section that the spammers lead the truthful users to follow them or correct them.

4.3.2 *Following* Effect

As we found in previous sections, the spammers' action activates following users (Chapter 4.2.2) and truthful users act together with the same opinion users in targeted products (Chapter 4.3.1). And now, we are going to see whether the truthful users behave in the incited direction by spammers. We use a same concept of window, but we gave a little of lags (10 points delay) between former-window and latter-window. The concentration of spams was still high right after the former-window due to the group attacks; accordingly, the behavior of the truthful users was buried. So with a delay, we show a relation between the *pReal* rate in latter-window and the *pSpam* rate in former-window (Figure 4.8(a)). For all windows with a same *pReal* rate, we use the average value of a corresponding *pSpam* rate.

Consequently, as a *pReal* rate increase, a *pSpam* rate also boosted

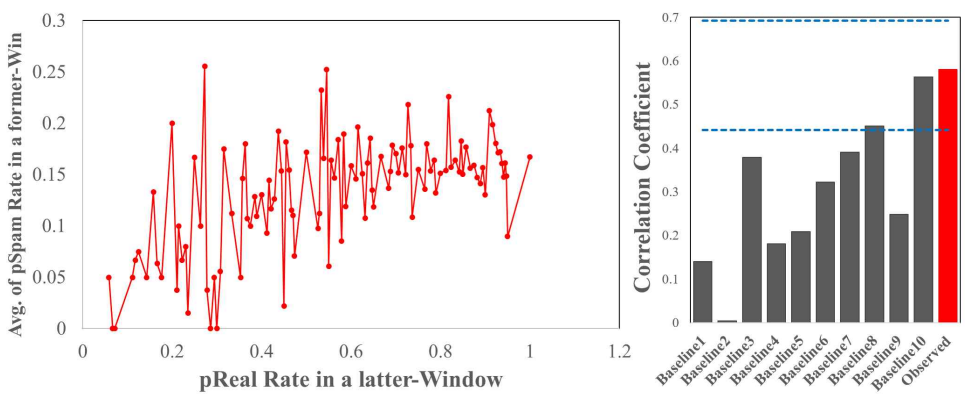


Figure 4.8: (a) A relation of pReal rate and pSpam rate (b) Pearson Correlation Coefficient of relation (a). Horizontal dashed blue lines indicate the confidence interval bound at 95% ($p<0.05$) confidence.

sequentially. The Pearson correlation coefficient (PCC) recorded 0.58 and we compared to baselines. We shuffled all the order of reviews except spam and calculated PCC about the relation. We repeated ten times, and all the result of them was lower than the real-world dataset (Figure 4.8(b)). The PCC value rejected null hypothesis ($p < 0.05$) indicating statistically significant correlation.

Through the correlation between these two type of reviews, it can be seen that the activity of *pSpam* reviews in the front window affects the *pReal* reviews' activity that follows. After all, truthful users are hooked to spammers and support their opinion as the owner wanted.

Following effect is an observation that polluted opinion from a spammer disturbs truthful users to evaluate frankly. It means that even if the system filters spammers, reliability isn't recovered because the followed opinion was already biased by the spammers.

4.3.3 *Correcting* Effect

We found that there is a phenomenon among truthful users which follows contrast theory after spammers' action. [25] have shown users gave a lower score than the quality if its' historical rating was high (as the users' expectation was high, and they would be disappointed with a real class). The meaning of "contrast" is an action that truthful users try to correct the opinion polluted by spammers and we call this behavior *"correcting"* effect. We investigate whether spammers trigger *correcting* behavior among truthful users.

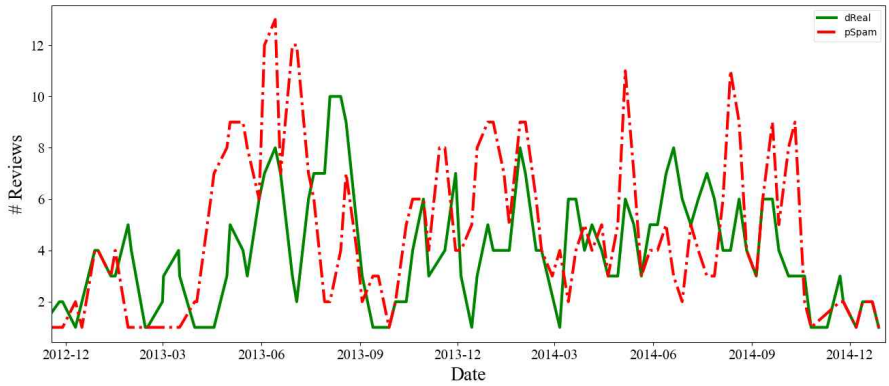


Figure 4.9: The number of (*dReal*, *pSpam*) type reviews for two weeks.

For example, a product gets higher average rating score than the actual quality instantaneously, by spammers' attacks, which increases the expectation of the quality to the following users who saw the exaggerated score. When the truthful user then confirms an actual quality of the product, he/she will be disappointed by the class that is much lower than expected and will emotionally leave a lower score on the platform. Figure 4.9 is an example of a real-world product which shows such phenomenon. During 03/2013 to 09/2013, after spam showed burst activity, dReal type reviews show similar patterns. Same in 09/2013 to 03/2014.

We can present the behavior generally in Figure 4.2. The probability of *dReal* appearance which has a higher value than the statistical baseline ($D^{(pSpam,dReal,k)} > 0$) was much more after spam occurred ($k > 0$) than before($k < 0$). The difference of 22 points out of 30 points over the baseline after spam happened, whereas, only 6 points before spam occurred. Among them(22 points, 6 points each), the average value was 9.48% for the latter and 5.11% for the former. And also, the difference increased overall ($PCC(k, D^{(pSpamdReal,k)}) = 0.73$,

After a terrible experience here last night and after reading many of the five-star Yelp reviews, I'm convinced that many of these reviews lavishing praise on the management and service (and eliciting comments from the manager such as "I love reading detailed reviews such as this describing every aspect of the restaurant."...hmmm.) were paid for by management. ··· PS: Bolstering my theory of positive review tampering, I'm sure that a few more 5-star reviews will pop up soon after mine goes up.

Figure 4.10: An example of *correction* review.

$-30 \leq k \leq 30$). To check whether the difference of an appearance probability before/after the spam is significant or not, we test the hypothesis at 0.05 significance level. Through the test, the observed data rejects the null hypothesis, which indicates there's a statistically significant difference between before and after. As a result, when spam is activated, not only the effect of *following* but also the phenomenon of reverting polluted opinion is happening.

Among *dReal* type reviews which try to recover the contamination, there are significant reviews which act resolutely. We call them ***correction* review**. Figure 4.10 shows an example of correction review. We pre-processed with specific words (e.g. 'hype' , 'overrated', and so on) that included most of the correction reviews in the 1-point rating for all products, and then manually found 67 correction reviews. One question is when does a correction review appear. The answer is in Figure 4.11.

We saw what happens right before a correction review appears. We set correction review as an anchor review and calculated the rate of spam appearance at former-window($W_f = 10$), a blue-bar in Figure 4.11(a). For comparison, we set each score rating(1 to 5) as an anchor point and calculate the same thing(Figure 4.11(a) sky-blue

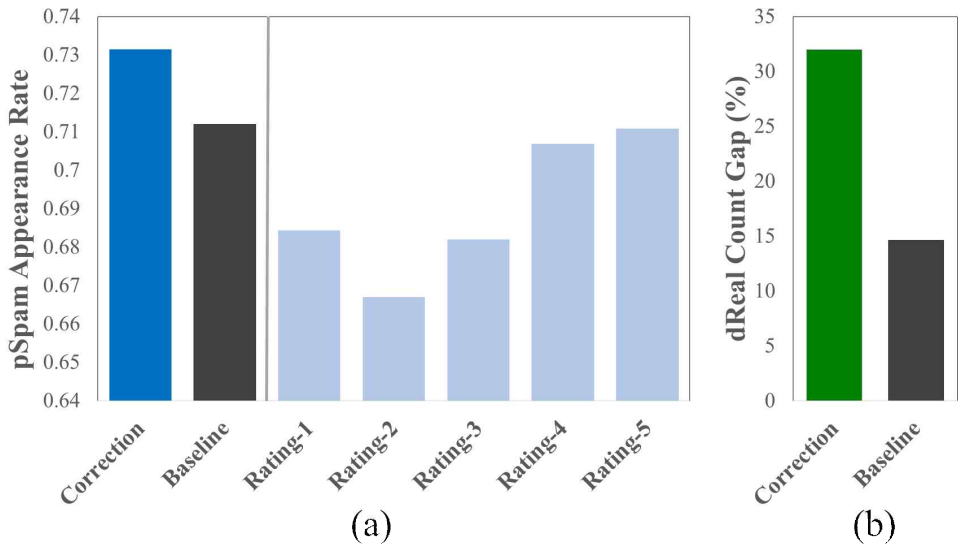


Figure 4.11: Phenomenon before the correction review. (a) pSpam appearance rate before each type of review. (b) dReal count gap before the correction review.

bar). And we also shuffle all the order of reviews except spam and then calculate on re-positioned correction reviews (Figure 4.11(a) gray bar). As shown in Figure 4.11(a), *pSpam* appearance rate right before the correction review is the highest. We observed that before the correction reviews, spam occurred more than others, and it signified a correction review reacts to spammers.

And we calculate a rate gap of *dReal* which represents the difference at before and after. If a correction review is $(r_c, d_c) \in R_p$, the former-window of it is $R_p(c - W_f)$ and the latter-window is $R_p(c+1)$.

An observed number of *dReal* type review in former-window and latter-window is:

$$C_p^D(c - W_f) = \left| \{ r | r \in R_p(c - W_f), r : dRealtype \} \right|$$

$$C_p^D(c+1) = \left| \{ r | r \in R_p(c+1), r : dReal\,type \} \right|$$

Then, the difference (%) of *dReal* count between the former and the latter is:

$$dReal\ Count\ Gap = \frac{C_p^D(c-ws) - C_p^D(c+1)}{C_p^D(c+1)} \times 100$$

Figure 4.11(b) stands for the observed gap. We also use a shuffled dataset as a baseline. Compare to the baseline, the difference shows more than double, and it means correction reviews occur behind other dReal reviews (same opinion as correction reviews) not ahead of them. We tested the hypothesis (i.e., dReal review count between before and after correction review is significantly different) at the 0.1 significance level and observed that null hypothesis of the t-test is rejected ($p < 0.1$). It indicates the difference is significant in statistically.

This behavior is similar to collective action [29]. Collective action in society is a theory that each user participates in an activity when the number of neighbors who have the same opinion of the users exceeds a certain threshold. Through our observation, when someone has to refute upon the contaminated score by spammers and the following effect, we pointed out that correction reviews occurred when the same opinion reviews appeared ahead of them; those reviews trigger correction reviews to act.

And now, we are going to compare two type of product, a product with correction and without correction. In chapter 4.3.1, we observed a group action in the same type reviews. Similarly, we introduced the **Index Gap Entropy (IGE)** to see if dReal reviews and pSpam reviews act in a group for each product. IGE means the entropy of

the index difference between two same adjacent type of reviews. In other words, the larger the entropy value, the more same type reviews stand together. We calculated IGE of the *dReal* and *pSpam* type for each product. IGE for *pSpam* can be calculated as follows:

$$H_p(X_S) = -\sum_{i=1}^{n} P(x_i)\log_b P(x_i)$$

where X_S is a discrete random variable for a product p, which contains a possible value $\{x_1,\ldots,x_n\}$, an index gap of two adjacent *pSpam* reviews. And IGE for *dReal* can be calculated in the same way:

$$H_p(Y_D) = -\sum_{i=1}^{n} P(y_i)\log_b P(y_i)$$

Y_D is also a discrete random variable for *dReal*s' index gap.

To get the entropy, we calculated all the possibility of all possible gap in a product, and we got a 2D point $(H_p(X_S), H_p(Y_D))$ for each product. If $H_p(X_S)$ is high, it means that *pSpam* type review occurs collectively. Same for $H_p(Y_D)$.

For an analytical baseline, we calculate IGE with Poisson approximation per each product. The probability of getting j index gap between two spam reviews in the product p is calculated as:

$$P(\hat{X}_S = j) = \frac{(\lambda_S)^j e^{-\lambda}}{j!}$$

$$\lambda_S = \frac{|R_p| + 1}{|\{r | r \in R_p, r : pSpam\ type\}| + 1}$$

where $\hat{X}_S \sim Poi(\lambda_S)$ is an index gap variable of two adjacent *pSpam* reviews in the product p, and λ_S is an expected index gap value of

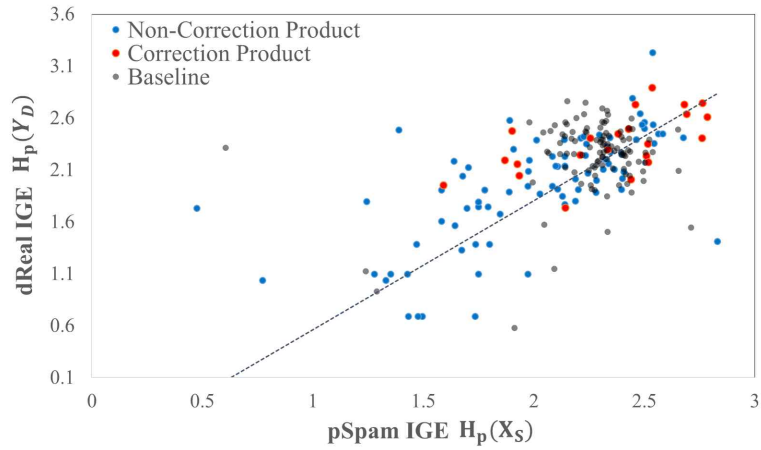


Figure 4.12: IGE graph

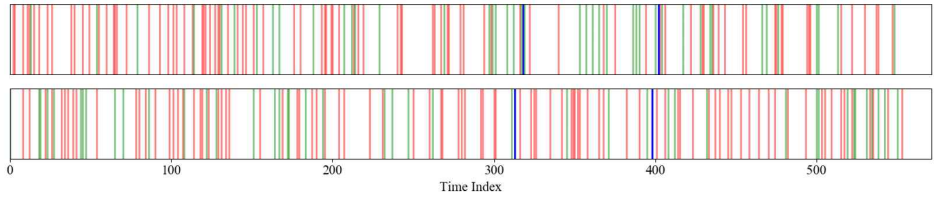


Figure 4.13: Observed(upper) and Shuffled(lower) series of review type. (white-pReal, red-pSpam, green-dReal, blue-correction)

two adjacent *pSpam* reviews. Same with *dReal*s' analytical baseline:

$$P(\hat{Y}_D = j) = \frac{(\lambda_D)^j e^{-\lambda}}{j!}$$

$$\lambda_D = \frac{|R_p| + 1}{|\{r | r \in R_p, r : dReal\ type\}| + 1}$$

$\hat{Y}_D \sim Poi(\lambda_D)$ is an index gap variable of two adjacent *dReal* reviews in the product p, and λ_D is an expected index gap value of two adjacent *dReal* reviews. With this approximated probability, we can calculate analytical IGE $(H_p(\hat{X}_S), H_p(\hat{X}_D))$ per each product.

Figure 4.12 shows *pSpam* IGE and *dReal* IGE per product and map to the graph. For the observed points (red, blue points), *pSpam* and *dReal* IGE show linear relation ($PCC = 0.68$), whereas, analytical baseline don't have any relationship. Also, the correction products (red points) have relatively higher value than the non-correction products'(blue points). Compare to the baseline, correction products which had a higher value in both *pSpam* and *dReal* (i.e. $H_p(X_S) > H_p(\hat{X}_S), H_p(X_D) > H_p(\hat{X}_D)$) was 59.09% of all, though, 12.94% for non-correction products.

Specifically, Figure 4.13 shows a real products' time series of review type which has a high IGE value. The upper one is an observed series, and the lower one is a shuffled data except for correction reviews. There was a distinct difference in their appearance pattern. Especially from 300 to 400 index in observed data, *dReal* type reviews show dense emergence and correction reviews appear right after them.

Through IGE, we observed products with correction reviews act more together with the same opinion than others.

Chapter V

Spam Attack Detection

In previous sections, we observed that both spam and *dReal* reviews occurred before correction review. If so, is it possible to predict whether spam attacks in a window using correction reviews' signal? In this section, we detect assaults from spammers using machine learning model only trained by correction reviews. We selected six features (4-5 review rate, 1-3 review rate, rating entropy, window period, average rating, variance) of a window for detection based on previous observations (chapter 4).

To check that correction review after spam attack is an indicative factor of spam attack, we compare the performance with a model trained by six times larger review set size. Concretely, we first trained models using only correction reviews (67 reviews with one-star rating), and compare with models trained by half of the one-star rating reviews (830 reviews). And then test both models with the other one-star rating reviews (not used in training). We use eight learning models, and Table 5.1 shows the performance. We used to adjust the concentration of spam from 5% to 20% as a criterion of attack.

As the table shows, correction review trained models showed a high level of performance, comparable to the performance of the models learned by half of the one-point reviews. All of the accuracies is

over than 0.7 and the F1-score is over than 0.8 when detecting more than 5% or 10% attacked window. Among eight models, support vector classifier (SVC) showed the best performance. The F1-score is zero because there is almost no attack window in which the concentration of spam is 20% or more.

The result shows that a small amount of correction review is an essential factor in the detection model. In other words, the correction review shows a clearer attack pattern than the other reviews, and a good detection model can be trained with a small amount of data. Also, windows' variance and average rating have the highest weight

Table 5.1: Attack classification result, A: Accuracy, AP: Average Precision, F1: F1-score for classification, KNN: K-nearest neighbors SVC: Support Vector Classifier, GP: Gaussian Process, DT: Decision Tree, RF: Random Forest, MLP: Multi-layer Perceptron, AB: AdaBoost Classifier, NB: Gaussian Naive Bayes.

		Spam ≥5%			Spam ≥10%			Spam ≥15%			Spam ≥20%		
	Classifier	A	AP	F1	A	AP	F1	A	AP	F1	A	AP	F1
Correction Trained	KNN	0.86	0.89	0.92	0.76	0.78	0.85	0.63	0.49	0.52	0.72	0.31	0.32
	SVC	0.86	0.87	0.93	0.76	0.76	0.86	0.62	0.48	0.2	0.77	0.37	0.29
	GP	0.86	0.86	0.93	0.72	0.76	0.82	0.6	0.48	0.53	0.73	0.27	0
	DT	0.8	0.89	0.88	0.73	0.76	0.83	0.64	0.5	0.53	0.67	0.34	0.46
	RF	0.79	0.9	0.87	0.7	0.77	0.81	0.62	0.48	0.43	0.75	0.35	0.37
	MLP	0.86	0.86	0.93	0.43	0.76	0.4	0.55	0.42	0.24	0.66	0.27	0
	AB	0.84	0.9	0.91	0.75	0.81	0.83	0.63	0.5	0.53	0.68	0.32	0.4
	NB	0.76	0.91	0.85	0.71	0.78	0.81	0.63	0.52	0.64	0.57	0.3	0.41
All Trained	KNN	0.84	0.9	0.91	0.74	0.8	0.82	0.68	0.54	0.58	0.71	0.34	0.42
	SVC	0.86	0.87	0.92	0.74	0.74	0.84	0.59	0.44	0.16	0.7	0.27	0
	GP	0.88	0.89	0.93	0.73	0.73	0.84	0.58	0.42	0	0.73	0.27	0
	DT	0.89	0.92	0.94	0.75	0.81	0.83	0.65	0.52	0.58	0.72	0.34	0.39
	RF	0.86	0.89	0.92	0.76	0.78	0.85	0.66	0.52	0.51	0.76	0.34	0.18
	MLP	0.88	0.87	0.93	0.72	0.73	0.83	0.57	0.42	0	0.76	0.34	0.18
	AB	0.89	0.92	0.94	0.72	0.77	0.82	0.7	0.57	0.6	0.76	0.37	0.39
	NB	0.84	0.91	0.9	0.72	0.77	0.82	0.62	0.51	0.65	0.56	0.33	0.48

of features in the SVC model. It is where the features we observed work (i.e., spam attack and *dReal* reviews work together in the front of correction review). Through this experiment, we effectively detected spammers' attack using significant attack signal.

Chapter VI

Conclusion

In this thesis, we analyzed users' behavior using real-world Yelp dataset. Especially, this is the first research which observed the behavior of truthful users triggered by spammers. Experiments using time-series analysis showed six novel observations. We proposed new methods to evaluate their behavior. All the findings are based on statistical or theoretical baseline, and we obtained significant patterns. We showed the same type of users herd each other, and this phenomenon was obviously found in targeted products. And spammers work after the appearance of low ratings, and their attack activated truthful users' movements. Two effects were observed responding to spammers: *following* and *correcting* effect. As spammers act strong, some truthful users are instigated to support spammers ("*following*"), in contrast, some users are triggered to fly in the face of polluted public opinion ("*correcting*"). According to the observations, we first revealed that the owner can get higher profits by employing spammers and thanks to the *following* effect; whereas, he/she can make an unexpected loss by the *correcting* effect which tries to counteract the impact of spammers. And we finally detected spammers' attack only using a small amount of correction reviews. The outcome was impressive since we only need a small set of reviews and common features for training.

Not only detecting the period of attack, but the observations we have made are also utilizable in various fields. We can predict user behavior after spammers' activity and can be used to provide or recommend a pure rating score, excluding users' comments biased by spammers. It is meaningful for providing reliable information and building trust between a store owner and customers.

BIBLIOGRAPHY

[1] Jindal, Nitin, and Bing Liu. “Opinion spam and analysis.” Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008.

[2] Spirin, Nikita, and Jiawei Han. “Survey on web spam detection: principles and algorithms.” ACM SIGKDD Explorations Newsletter 13.2 (2012): 50-64.

[3] Ott, Myle, Claire Cardie, and Jeff Hancock. "Estimating the prevalence of deception in online review communities." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[4] Ott, Myle, et al. "Finding deceptive opinion spam by any stretch of the imagination." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[5] Lim, Ee-Peng, et al. "Detecting product review spammers using rating behaviors." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[6] Hooi, Bryan, et al. "Birdnest: Bayesian inference for ratings-fraud detection." Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial

and Applied Mathematics, 2016.

[7] Mukherjee, Arjun, et al. "Spotting opinion spammers using behavioral footprints." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

[8] KC, Santosh, and Arjun Mukherjee. "On the temporal dynamics of opinion spamming: Case studies on yelp." Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.

[9] Fei, Geli, et al. "Exploiting Burstiness in Reviews for Review Spammer Detection." Icwsm 13 (2013): 175-184.

[10] Akoglu, Leman, Rishi Chandy, and Christos Faloutsos. "Opinion Fraud Detection in Online Reviews by Network Effects." ICWSM 13 (2013): 2-11.

[11] Rayana, Shebuti, and Leman Akoglu. "Collective opinion spam detection: Bridging review networks and metadata." Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining. ACM, 2015.

[12] Shin, Kijung, Bryan Hooi, and Christos Faloutsos. "M-zoom: Fast dense-block detection in tensors with quality guarantees." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2016.

[13] Shin, Kijung, et al. "D-cube: Dense-block detection in terabyte-scale tensors." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining.

ACM, 2017.

[14] Liu, Shenghua, Bryan Hooi, and Christos Faloutsos. "HoloScope: Topology-and-Spike Aware Fraud Detection." Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017.

[15] Ye, Junting, Santhosh Kumar, and Leman Akoglu. "Temporal Opinion Spam Detection by Multivariate Indicative Signals." ICWSM. 2016.

[16] Hooi, Bryan, et al. "Fraudar: Bounding graph fraud in the face of camouflage." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[17] Kaghazgaran, Parisa, James Caverlee, and Anna Squicciarini. "Combating Crowdsourced Review Manipulators: A Neighborhood-Based Approach.", Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018.

[18] Li, Huayi, et al. "Bimodal distribution and co-bursting in review spam detection." Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017.

[19] Jiang, Meng, et al. "Spotting suspicious behaviors in multimodal data: A general metric and algorithms." IEEE Transactions on Knowledge and Data Engineering 28.8 (2016): 2187-2200.

[20] Baddeley, Michelle. "Herding, social influence and economic decision-making: socio-psychological and neuroscientific

analyses." Philosophical Transactions of the Royal Society of London B: Biological Sciences 365.1538 (2010): 281-290.

[21] Guille, Adrien, et al. "Information diffusion in online social networks: A survey." ACM Sigmod Record 42.2 (2013): 17-28.

[22] Lee, Young-Jin, Kartik Hosanagar, and Yong Tan. "Do I follow my friends or the crowd? Information cascades in online movie ratings." Management Science 61.9 (2015): 2241-2258.

[23] Romero, Daniel M., Katharina Reinecke, and Lionel P. Robert Jr. "The Influence of Early Respondents: Information Cascade Effects in Online Event Scheduling." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017.

[24] Lederrey, Gael, and Robert West. "When Sheep Shop: Measuring Herding Effects in Product Ratings with Natural Experiments." Proceedings of the 27th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2018.

[25] Zhang, Xiaoying, Junzhou Zhao, and John Lui. "Modeling the Assimilation-Contrast Effects in Online Product Rating Systems: Debiasing and Recommendations." Proceedings of the Eleventh ACM Conference on Recommender Systems. ACM, 2017.

[26] Li, Huayi, et al. "Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns." ICWSM. 2015.

[27] Mukherjee, Arjun, et al. "What yelp fake review filter might be doing?." ICWSM. 2013.

[28] Xu, Chang, Jie Zhang, and Zhu Sun. "Online reputation fraud campaign detection in user ratings." Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press, 2017.

[29] Easley, David, and Jon Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, 2010.

요 약

의견 스패머에 대응하는 사용자 행동 분석: 동조 및 자정 효과

박지현

컴퓨터공학부

서울대학교 대학원

온라인 리뷰 플랫폼에서 의견 스팸은 여전히 해결되지 못한 문제 중 하나이다. 의견 스팸은 고도화된 스팸 기법으로 인하여 사람의 눈으로도 스팸 여부를 판단하기 어렵다. 특히 기존의 다른 종류의 스팸은 언어 정보로도 쉽게 스팸 여부를 판단할 수 있으나, 의견 스팸은 이러한 강력한 탐지 요소를 사용하기 힘들다는 점이 의견 스팸 탐지의 가장 큰 문제이다. 본 논문에서는 기존 논문에서 다루지 않았던 새로운 접근법으로 의견 스팸의 영향력과 이에 반응하는 사용자의 행동을 분석하였다. 스패머의 공격 패턴과 공격 이후 사용자의 활동량이 늘어남을 발견하였다. 일반적으로 플랫폼에서 사용자는 이전 사용자의 의견에 따라가는 현상이 보였으며, 이 현상은 스팸이 활발히 활동하고 있는 제품에서 더 뚜렷이 발견됐다. 또한, 스패머의 공격 이후 그들의 의견에 선동되어 따라가는 사용자들과 오히려 스패머에 의해 오염된 사회망을 정화하려는 사용자들

의 행동이 나타났다. Yelp 데이터를 사용하여 스패머 발생 시점 주변에서 나타나는 일시적 현상을 분석하였으며, 경험적 및 통계적 확률로 현상 신호를 나타내었다. 우리는 최종적으로 본 연구에서 밝힌 특징을 활용하여 스패머의 공격 구간을 탐지하였으며, 좋은 성능을 보임으로써 특징이 스팸을 탐지하는 데 효과적임을 보였다. 의견 스팸에 반응한 사용자의 행동을 분석한 최초의 논문이며, 우리는 스패머의 공격 전략, 효과, 그리고 스패머에 대응하는 일반 사용자들의 행동 패턴(동조 및 자정 효과)을 밝혔다.

주요어: 의견 스팸, 리뷰 스팸, 사용자 행동 분석, 의견 스팸 탐지, 사회적 영향, 자정 효과, 동조 효과

학번: 2016-21206