EQUAL RATING OPPORTUNITY ANALYSIS FOR DETECTING REVIEW MANIPULATION

Yongbo Zeng, Yihai Zhu, Yan (Lindsay) Sun
University of Rhode Island, Kingston, RI
Email: {yongbozeng, yhzhu, yansun}@ele.uri.edu

***Abstract*—Online review plays an important role when people are making decisions to purchase a product or service. It is shown that sellers can benefit from boosting their product review or downgrading their competitors' product review. Dishonest behavior on reviews can seriously affect both buyers and sellers. In this paper, we introduce a novel angle to detect dishonest reviews, called Equal Rating Opportunity (ERO) evaluation. The proposed ERO evaluation can detect embedded manipulation signals based on limited amount of data. Experiments based on real data are conducted. Four highly problematic products are successfully detected from 303 products.**

***Index Terms*—Reputation system, Trust, Security**

I. INTRODUCTION

Online review system, also referred to as *online reputation system*, allows users to post reviews for products or services, and aggregate these reviews to a reputation score that indicates the user satisfaction estimation. One example of such reputations scores is the number of stars in Amazon. Online reputation systems can help people evaluate the quality of products or services before transactions, and hence greatly reduce the risk of online shopping. Online reviews are posted by people who have experiences of using the products or services. One review usually consists of a rating score indicating the user satisfaction, and a piece of comment describing the experience of using the product or service. Examples of online review include Amazon customer review, Yelp review, online hotel review, etc. People are becoming more and more relying on online reviews when evaluating the quality of products, hotels, restaurants, and vacation packages before paying for them.

However, online reviews may be manipulated, since there are huge profits of online markets [1] and the purchasing decisions can be misled by dishonest reviews. It is reported that sellers at the online marketplace boost their reputation by trading with collaborators [2], and firms post biased reviews to praise their own products or bad-mouth the competitors' products [3]. Review manipulation can inflate or deflate products' reputation scores, crash users' confidence in online reputation systems, and eventually undermine reputation-centric online businesses leading to economic loss. Furthermore, there are some situations, in which the review manipulation is even more damaging. For example, *Black Friday* shoppers heavily relies on online reviews, because they have to make rush decisions for the products they are not familiar with in order to take advantage of these quickly expiring 'unusual' discount. Another example is online reputation of hotels and restaurants. The consumers, who are misled by manipulated hotel ratings, cannot be easily refunded after they purchase these services.

In the literature, researchers propose methods to protect reputation systems from several angles, such as 1) increasing the cost of acquiring multiple user IDs [4], 2) endogenous discounting of dishonest reviews by analyzing the statistic features of the reviews [5], 3) exogenous discounting of dishonest ratings by introducing reputation evaluation of users [5]–[7], and 4) studying correlation between users and reviews to detect dishonest reviews [8], [9]. In this paper, we propose a new metric to detect dishonest reviews, called Equal Rating Opportunity (ERO) evaluation. This method roughly belongs to category 2 and 4.

The proposed ERO evaluation has two major advantages.

- It does not require the cooperation of online reputation system owner. In particular, many existing algorithms need to use a large amount of data, which makes them impractical unless the reputation system owner (e.g. Amazon and Yelp) implements these algorithms. Our ERO approach, however, can be implemented by a third party, who only needs to crawl a very small amount of data to train the detection parameters and perform the ERO evaluation. Therefore, The proposed method is a low-cost solution, yields independent opinions, and leads to practical implementations.
- ERO evaluation is a new direction for detecting review manipulation. It is compatible with most of existing algorithms, which exam ratings and reviewers from more traditional angles. ERO has a potential to find the manipulation signals, which were previously missed by existing approaches.

In this paper, related work is discussed in Section II. The ERO analysis and rating consistency detection algorithm are presented in Section IV and Section III respectively. We then conducted real data testing to evaluate the features and performance of ERO analysis. Totally 303 products in 4 categories are analyzed. For example, among 84 'Toys & Games' products, our methods detect 4 problematic products. Two products are confirmed to have review manipulation problem, the other two are highly suspicious too. These experiments and results are shown in Section V.

II. RELATED WORK

In order to protect online reputation systems, researchers propose many protection schemes that can be roughly put into 4 categories. The **first** category is increasing the cost of acquiring multiple user IDs by binding user IDs with IP addresses [4]. The **second** category is endogenous discounting of dishonest reviews [5]. Dishonest ratings are directly differentiated from normal ratings based on the statistic features of the rating values. In a Beta-function based approach [10], a user is determined as a malicious user if the estimated reputation of the product rated by him/her lies outside q and $1-q$ quantile of his/her underlying rating distribution. An entropy based approach is proposed in [11]. The **third** category is exogenous discounting of dishonest ratings. Users are assigned trust scores based on their review history, and the quality of their reviews are discounted according to their trust scores. In [7], a user's trust is obtained by cumulating his/her neighbors' beliefs through belief theory. The **fourth** category is studying correlation between users and reviews to detect dishonest ratings [8], [9]. The proposed scheme has both category 2 and category 4 features, and the detection algorithm is from a new angle.

Many research results did not turn into practical systems. This is probably because of the potential liability concerns of major e-commerce companies, as well as the gap between research and practical constraints. Without the support from the e-commerce companies (i.e., reputation system owners), the algorithms can only rely on a limited amount of data. This is one of the major hurdles. Currently, there are only a few existing online systems providing review analysis services. For example, there is a website called "ReviewPro" [12], whose major business is to provide professional suggestions to hotel owners. By analyzing the customers' reviews on a hotel, ReviewPro can provide analytical reports with "strategies" to climb TripAdvisor rankings and earn 5-star reviews. Another practical system is "TrustYou" [13], which provides review analysis services on hotels. For hotel owners, it provides service to market the reputation and increase businesses. For individual users, it provides service to analyze the hotel's quality, by summarizing online reviews and generating a trust score for the hotel. What we propose in this work is fundamentally different from these existing services. First, our work focus on detecting review manipulation, instead of finding patterns for reputation promotion purpose. Second, our work can provide on-demand real time service, whereas ReviewPro and TrustYou can only offer analysis of a pre-determined list of hotels. That is, our algorithm is so effective that it can detects manipulation signals based on the small amount of data crawled in real time.

III. Consistency Analysis

Consistency analysis is a popular and effective review manipulation detector, which detects if the rating scores of a product are inconsistent with time. This type of detector is based on the fact that in order to perform an effective manipulation, the dishonest ratings must cause large enough change in the average rating. In the literature, there are several approaches to detect the variation of average ratings. In this paper, we adopt the one proposed in [14] called CUSUM, as an pre-process for ERO analysis. In this section, we will briefly describe this approach.

We define the notations used in this paper as follows.

- p_i is the product with product id i.
- $r_{i,n}, n = 1, 2, \ldots, N$ is the nth review of p_i, where the reviews are sorted by posting time from old to new.
- $r_{i,n} = \{x_{i,n}, f^{(1)}_{i,n}, f^{(2)}_{i,n}, \ldots, f^{(K)}_{i,n}\}, k = 1, 2, \ldots, K$, where $x_{i,n}$ is the rating, and $f^{(K)}_{i,n}$ is the observation of the kth ERO feature.
- We also use vector notations. $\mathbf{r_i}$ is the review vector, $\mathbf{x_i}$ is the rating vector, and $\mathbf{f_i^{(k)}}$ is the ERO feature observation vector.

Let μ_i be the true rating of p_i, if $\bar{\mathbf{x}}_\mathbf{i} > \mu_i + \nu$ or $\bar{\mathbf{x}}_\mathbf{i} < \mu_i - \nu$, change of average rating is observed, where $\bar{\mathbf{x}}_\mathbf{i} = \frac{1}{N}\sum_{n=1}^{N} x_{i,n}$ is the average rating. The detection functions are:

$$\begin{cases} g^+_{i,n} = \max(g^+_{i,n-1} + x_{i,n} - \mu_i - \nu/2,\ 0) \\ g^-_{i,n} = \max(g^-_{i,n-1} - x_{i,n} + \mu_i - \nu/2,\ 0) \end{cases} \tag{1}$$

where $g^+_{i,n}$ indicates the positive changes, $g^-_{i,n}$ indicates the negative changes, and $g^+_{i,0} = 0$, $g^-_{i,0} = 0$ for initialization. *Rating inconsistency* is observed when g^+_n or g^-_n exceeds the threshold $\bar{h}$. Another metric, Percentage of Change Interval (PCI), is defined as

$$PCI(\bar{h}) = \frac{N_D}{N} \tag{2}$$

where N_D is the number of $g^+_{i,n}$ or $g^-_{i,n}$ points exceeds $\bar{h}$. Obviously, PCI value depends on the selection of threshold.

It is pointed that a uniform threshold for all products is not applicable, and heterogeneous thresholds should be used [14]. Briefly speaking, the threshold depends on $PCI(\bar{h}_0)$, where $\bar{h}_0$ is a predefined minimum threshold. Smaller $PCI(\bar{h}_0)$ gives smaller threshold, and vice versa. In this paper, we use the PCI value $PCI(\bar{h})$ as the evaluation of rating inconsistency. $PCI(\bar{h}) = 0$ means the ratings are consistent, $PCI(\bar{h}) = 1$ means strong inconsistency is observed. More detailed discussion on PCI can be found in [14]

IV. ERO Principle and ERO Analysis

A. ERO Principle

The consistency detection, which is based on the statistics of ratings, can only be used to find products that are suspected to be under review manipulation, but is lack of the capability to accurately detect such manipulation. This is because the normal rating can change without any manipulation. For example, when a restaurant changes the chief, a seller changes his/her attitude toward consumer complaints, and the manufacturer fixes a defect of the product, the ratings for the restaurant/seller/product could change. The rating is also related to price. Consumers tend to be more tolerant if they purchase deeply discounted products. If the price changes dramatically, the ratings may change. Therefore, after the consistency analysis gives us a set of suspicious products, we must apply a more informative analysis to confirm the review manipulation.

We are inspired by the *Equal Employment Opportunity Policy*, adopted by many employers. One example of such policy statement is as follows.

"All employment decisions at the company are based on business needs, job requirements and individual qualifications, without regard to race, color, religion or belief, national, $\cdots$."

We introduce an **Equal Rating Opportunity (ERO) Principle**, as follow.

ERO Principle. *"The normal ratings should be primarily based on the quality of the product or service, without regard to whether the review is posted on weekdays or weekend, posted during daytime or night time, long or short, $\cdots$."*

This principle is based on the idea that the dishonest raters may maintain a lot of user accounts and review templates, and if they perform review manipulation on a product, it can change the distribution of some features. Such change is called *manipulation signal*. For example, if the dishonest reviews are posted within a short time, such as on Sunday, it will increase the correlation between date and rating. Even if the dishonest raters randomly select a user account from a large account pool and post the dishonest review on a random day, this behavior can also change the distribution of other features. We suggest to use multiple ERO features, as described in Section V-B.

B. ERO Feature

Review features that apply to ERO principle are referred to as **ERO features**. Of course, there are some review features not applicable for ERO principle. For example, if a product is more favorable in the east coast than in the west coast, then the east coast users give higher ratings than the west coast users. On the other hand, we can find some factors that can be used as ERO features, such as "day of a week" and "time in a day". Based on the ERO principle, the "day of a week" feature of the 1-star reviews and the "day of a week" feature of 5-star reviews should be similar. If not, ERO principle is not satisfied and we argue that the product is highly suspicious to be a review manipulation victim.

C. ERO Analysis Process

In this subsection, the proposed ERO analysis is presented. The ERO evaluation takes the rating score vector $\mathbf{x_i}$ and the ERO feature observation vector $\mathbf{f_i^{(k)}}$ as inputs. Let $E_i^{(k)}$ denote the ERO value for $\mathbf{f_i^{(k)}}$. We propose to use the Pearson correlation coefficient as the ERO metric.

$$E_i^{(k)} = \frac{cov(\mathbf{x_i}, \mathbf{f_i^{(k)}})}{\sigma_{\mathbf{x_i}}, \sigma_{\mathbf{f_i^{(k)}}}} \tag{3}$$

where $cov(\mathbf{x_i}, \mathbf{f_i^{(k)}})$ is the covariance of $\mathbf{x_i}$ and $\mathbf{f_i^{(k)}}$, $\sigma_{\mathbf{x_i}}$ is the standard deviation of $\mathbf{x_i}$, and $\sigma_{\mathbf{f_i^{(k)}}}$ is the standard deviation of $\mathbf{f_i^{(k)}}$.

As we discussed in Section V-B, different categories can use different ERO features and thresholds, and for a specific category, one ERO feature will provide one ERO detector. The detector for $\mathbf{f_i^{(k)}}$ is trained as follows. M products are sampled from the market, and the ERO values $E_{d1}^{(k)}, E_{d2}^{(k)}, \ldots, E_{dM}^{(k)}$ are calculated. Next, we calculate the first quartile denoted by q_1 and the third quartile denoted by q_3. Finally, the detection thresholds are calculated as

$$\begin{cases} \theta_{upper} = q_3 + w(q_3 - q_1) \\ \theta_{lower} = q_1 - w(q_3 - q_1) \end{cases} \tag{4}$$

TABLE I: Summary of dataset

Category	# prod.	Avg. # reviews	Avg. rating
Electronics	44	1204	3.90
Home & Kitchen	89	808	4.08
Toys & Games	84	672	4.23
Video Games	86	1053	4.24

where w is used to determine the detector sensitivity. In the experiments , $w = 1.5$. $\theta_{upper} - \theta_{lower}$ is called the length of safe range. During the detection stage, if $E_i^{(k)} < \theta_{lower}$ or $E_i^{(k)} > \theta_{upper}$, then it is concluded that ERO detector of $\mathbf{f_i^{(k)}}$ detectes p_i to be highly manipulation suspicious.

For a given product p_i, there will be K detection results. In this paper, we consider the product to be suspicious when at least one detector returns 'highly manipulation suspicious'. In practice, the fusion method can be more complicated depending on the number of ERO feature and the detection rate requirement.

V. Experiment and Results

In this section, we apply the ERO detection on a real dataset crawled from Amazon.

A. Dataset

We develop web crawlers to gather real data from Amazon. The crawlers are implemented using Python [15]. The dataset we use in the experiments is collected as follows. We randomly select products under several guidelines. 1) The number of reviews is greater than 50. 2) The product is from category set {'Electronics', 'Home & Kitchen', 'Toys & Games', 'Video Games'}. 3) The average rating value is between 2.5 and 4.8. The summary of the dataset used in this section is listed in Table I.

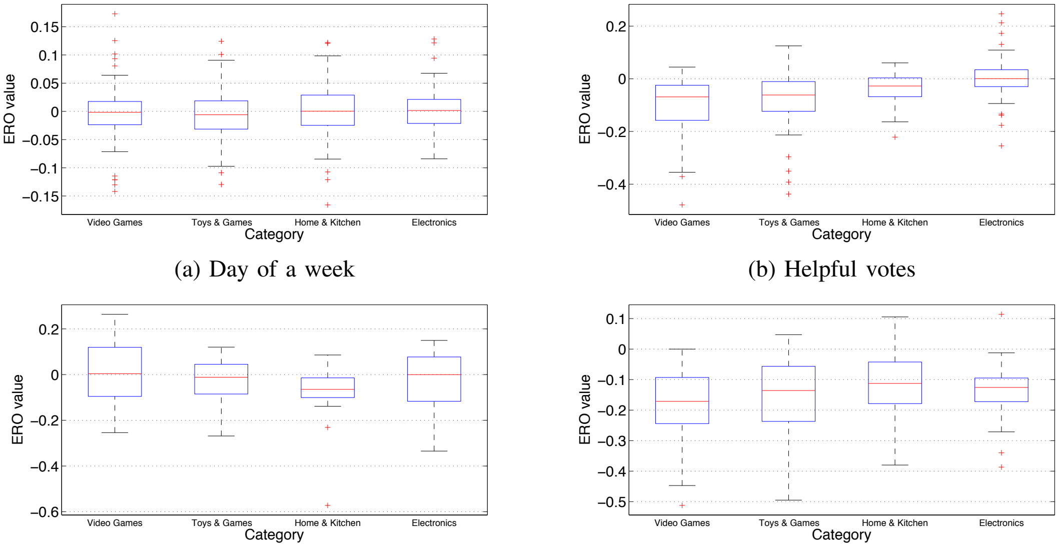
B. ERO Feature Selection

The normal rating values should be primarily based on the quality of the product or services, without regard to certain review features. ERO principle tells us that given the observations of an ERO feature, it is impossible to predict the rating values. It is important to point out that not all the review features are suitable for ERO analysis.

In this section, we conduct experiments to study the suitability of review features for ERO analysis. For a given category and a review feature, we run the ERO evaluation process and get one ERO value per product. There are total 4 features investigated, including 'day of a week', 'helpful votes', 'comment length' and 'customer ranking'. We plot the results in Figure 1, in which the x-axis is category, the y-axis is ERO evaluation, the red bar is the median and the balck bars are the detection thresholds. We decide whether a feature is suitable for ERO analysis if 1) the median is close to 0 (within $0 - e_0$ and $0 + e_0$), and 2) the length of safe range (defined in Section IV-C) is less than e_1. We use empirical values for e_0 and e_1, and in our experiment $e_0 = 0.02$ and $e_1 = 0.1$. The suitability of review features for each category are presented in Table II.

TABLE II: ERO feature selection

Feature \ Category	Electronics	Home & Kitchen	Toys & Games	Video Games
day of week	✓	✓	✓	✓
helpful votes	✓			
comment character length				
customer ranking	✓		✓	✓



(a) Day of a week

(b) Helpful votes

(c) Customer ranking

(d) Comment length

Fig. 1: Statistics of ERO values using different review features

TABLE III: Information of suspicious products

Product	p_a	p_b	p_c	p_d
# of reviews	189	194	232	180
Avg rating	**4.1**	**3.7**	**3.2**	**4.3**
$E^{(DOW)}$	-0.1089	0.1241	0.1009	-0.1295
CUSUM PCI	0.0212	0.1804	0.0862	0.0556
Other sites	Target **3.5** Newegg **3**	Tower Hobbies **4.8** RobotShop **4.8**	Google **3.6**	Target **4.2** Walmart **4.5**

For example, in the video game category, people tend to vote very high rating score or very low ratings scores as helpful, and rarely vote neutral ratings as helpful (or unhelpful). Then, the EOR value of ‘helpful votes’ won’t be useful for the video game Category. For electronics, however, this phenomenon does not exist. Then, the EOR value of ‘helpful votes’ is useful for the electronics category.

C. ERO Analysis

After we identify ERO features for different categories, we are able to establish the detection thresholds as equation (4).

In this section, we demonstrate the detection results of ‘Toys & Games’ category with ERO feature ‘day of a week’ (DOW). In Figure 1a, we observe 4 products out of the DOW detector’s safe range. We look at those 4 products and list the information in Table III. We also calculate the CUSUM PCI values for those 4 products, and there is only one product gives relatively high PCI value. However, it is extremely difficult to tell whether the products are manipulated or not. Therefore, we compare their rating scores to these on other websites. We confirm that product p_a and p_b are surely suspicious. For the other two products, although we cannot say for sure they are suspicious, we can say that the rating value is highly correlated with the day when the rating is given. For p_c, the ratings given on Sunday through Tuesday are much lower than the ratings given on Wednesday through Friday. For p_d, the ratings given on Friday and Saturday are much lower than the other days. Violating the ERO principle itself is considered as a signal of manipulation.

VI. Conclusion

The ERO principle, which is a new angle to detect review manipulation, has been proposed in this paper. Together with the consistency analysis as an pre-process, the manipulation signal detected by ERO is very different from the traditional approaches. Furthermore, it needs very limited data to set up detection thresholds, and only require the ratings for a particular product in order to determine whether this product is under review manipulation. Therefore, it can be performed in real-time. Experiments based on real data are conducted, the detected problematic products are confirmed to have manipulation problem. In the future, we look forward to performing testing with more real data and real user involvement.

References

[1] *Final Pre-Christmas Push Propels U.S. Online Holiday Season Spending Through December 26 to Record $30.8 Billion.* comScore, 2010. [Online]. Available: http://ir.comscore.com/releasedetail.cfm?ReleaseID=539354

[2] J. Morgan and J. Brown, “Reputation in online auctions: The market for trust,” *California Management Review*, vol. 49, no. 1, pp. 61–81, 2006.

[3] A. HARMON, *Amazon glitch unmasks war of reviewers*. The New York Times, February 14, 2004. [Online]. Available: http://www.nytimes.com/2004/02/14/us/amazon-glitch-unmasks-war-of-reviewers.html

[4] M. Abadi, M. Burrows, B. Lampson, and G. Plotkin, "A calculus for access control in distributed systems," *ACM Trans. Program. Lang. Syst.*, vol. 15, no. 4, pp. 706–734, Sep. 1993. [Online]. Available: http://doi.acm.org/10.1145/155183.155225

[5] A. Jsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Systems*, vol. 43, no. 2, pp. 618 – 644, 2007, emerging Issues in Collaborative Commerce. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923605000849

[6] P. Laureti, L. Moret, Y.-C. Zhang, and Y.-K. Yu, "Information filtering via iterative refinement," *EPL (Europhysics Letters)*, vol. 75, no. 6, p. 1006, 2006.

[7] B. Yu and M. P. Singh, "An evidential model of distributed reputation management," in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, ser. AAMAS '02. New York, NY, USA: ACM, 2002, pp. 294–301.

[8] Y. Liu and Y. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, Aug 2010, pp. 65–72.

[9] C. Dellarocas, "Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior," in *Proceedings of the 2Nd ACM Conference on Electronic Commerce*, ser. EC '00. New York, NY, USA: ACM, 2000, pp. 150–157. [Online]. Available: http://doi.acm.org/10.1145/352871.352889

[10] A. Jsang and R. Ismail, "The beta reputation system," in *Proceedings of the 15th bled electronic commerce conference*, 2002, pp. 41–55.

[11] W. Jianshu, M. Chunyan, and G. Angela, "An entropy-based approach to protecting rating systems from unfair testimonies," *IEICE TRANSACTIONS on Information and Systems*, vol. 89, no. 9, pp. 2502–2511, 2006.

[12] *ReviewPro*. [Online]. Available: http://www.reviewpro.com/

[13] *TrustYou*. [Online]. Available: http://www.trustyou.com/

[14] Y. Liu, Y. Sun, and T. Yu, "Defending multiple-user-multiple-target attacks in online reputation systems," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, Oct 2011, pp. 425–434.

[15] Y. Zeng, Y. Zhu, and Y. Sun, "Reviewsec: Security analysis tool for online reviews," *STCSN E-Letter*, vol. 2, no. 4, 2014.