IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY VOL:8 NO:6 YEAR 2013

Securing Online Reputation Systems through Dempster-Shafer Theory based Trust Model

Yuhong Liu[1], Yan (Lindsay) Sun[1], Siyuan Liu[2], and Alex C. Kot[2]

[1]Department of Electrical and Computer Engineering University of Rhode Island, Kingston, RI 02881
[2]School of Electrical and Electrinic Engineering, Nanyang Technological University, Singapore
Emails:{yuhong, yansun}@ele.uri.edu, {lius0036, eackot}@ntu.edu.sg

***Abstract*—With the rapid development of online reputation systems, manipulations against such systems are evolving quickly. In this paper, we propose a Dempster-Shafer theory based trust scheme to protect reputation systems. When tested against real user attack data collected from a cyber competition, the proposed scheme has achieved a very good performance in terms of accurately identifying malicious users. It also demonstrates a great potential to effectively remove dishonest ratings and keep the online reputation system a secure and fair marketplace.**

I. Introduction

Online reputation systems are playing increasingly important roles in influencing people's online purchasing/downloading decisions. Meanwhile, manipulations against such systems which overly inflate or deflate reputation scores of online items are evolving rapidly. For example, for just $9.99, a video on YouTube could receive 30 "I like" ratings or 30 real user comments provided by "IncreaseYouTubeViews.com". Without proper defense schemes, attacks against reputation systems can overly inflate or deflate item reputation scores, crash users' confidence in online reputation systems, and lead to economic loss.

The existing defense mechanisms protect reputation systems from several angles. First, *limit the maximum number of ratings* each user could provide within a certain time duration [1]. Second, *increasing the cost of acquiring multiple user IDs* by binding identity with IP address [2] or using network coordinates to detect sybil attacks [3]. Third, *investigating rating distributions* through statistical techniques, such as beta-function based approach [4], entropy based scheme [5], and Bayesian model based method [6]. Fourth, *investigating users' rating behaviors* by building user trust, such as a personalized trust model [7], an iteration refinement approach [8] which evaluates a user's "judging power" as the inverse of this user's rating variance.

In this paper, we propose a Dempster-shafer theory based trust model to identify *malicious users* who insert unfair ratings to mislead items' reputation scores. Specifically, we investigate time domain rating information for analyzing user behavior anomaly and based on this, build up the Dempster-Shafer theory based trust model to further identify malicious users. When tested against real user attack data collected from a cyber competition, it has demonstrated a good performance in terms of identifying malicious users.

II. Proposed Scheme

In this section, we will discuss details of the components of the proposed scheme.

A. Temporal Analysis - Change Detector

In the temporal analysis, we organize the ratings to a given item as a sequence in the descending order according to the time when they are provided. In many practical reputation systems, items have intrinsic and stable quality, which should be reflected in the distribution of normal ratings. Therefore, rapid changes can serve as indicators of anomaly. We propose a revised CUSUM detector [9] as the anomaly detector, which reliably detects changes occurring in the rating sequences of an online item.

Specifically, we want to determine whether a parameter θ in a probability density function (PDF) has changed. That is, to determine between two hypothesis: $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. Let p_{θ_0} and p_{θ_1} denote the PDF before and after the change, respectively. Let y_k denote the k^{th} sample of the data sequence (i.e. rating sequence). The change decision function is

$$g_k = max(g_{k-1} + \ln \frac{p_{\theta_1}(y_k)}{p_{\theta_0}(y_k)}, 0) \ , \tag{1}$$

$$t_a = min\{k : g_k \geq \overline{h}\} \ , \tag{2}$$

where $\overline{h}$ is the change detection threshold. Here, t_a is called *stopping time*, the time when the detector identifies a change and raises an alarm.

When p_{θ_0} is a Gaussian process with mean μ_0, p_{θ_1} is a Gaussian process with mean μ_1, and both have variance σ^2, equation (1) detects the mean change and becomes

$$g_k = max(g_{k-1} + (y_k - \mu_0 - \frac{\mu_1 - \mu_0}{2}), 0) \ . \tag{3}$$

The proposed detector is used to (a) detect whether changes occur in the ratings of an item and (b) estimate the time when attacks are suspected (i.e. which rating is suspicious).

B. Trust Model based on the Dempster-Shafer Theory

Based on the anomaly detection results, we further evaluate users' trust values in this section. In most trust models, users' trust values are determined only by their good and bad behaviors. However, it is not sufficient. Consider two trust calculation scenarios. First, user A has conducted 5 good behaviors and 5 bad behaviors. Second, user B is a new coming user and has no behavior history. In several trust models [4] [6], both of their trust values will be calculated as 0.5, although we are more confident in user A's trust value. To differentiate these two cases, the concept of behavior uncertainty is introduced by the Dempster-Shafer theory, to represent the degree of the ignorance of behavior history. In this work, we adopt the behavior uncertainty by proposing a trust model based on the Dempster-Shafer theory.

1) **The Dempster-Shafer Theory:** a framework for combining evidence from different sources to achieve a degree of belief [10].

Suppose two events $\{a, b\}$ are under consideration, where a = good behavior and b = bad behavior, and a subject is observed to perform good behaviors for r times, and perform bad behaviors for s times.

$$\begin{cases} B_g = \frac{r}{r+s+2}, \\ B_b = \frac{s}{r+s+2}, \\ u = \frac{2}{r+s+2}, \end{cases}$$

where B_g is the belief that the proposition that the subject will perform good behavior is true, B_b is the belief that the proposition that the subject will perform bad behavior is true, and u is the uncertainty.

2) **Trust Model Using the Dempster-Shafer Theory:** Based on the anomaly detector, for each given item, we could determine which ratings are suspicious. We then define a user's **behavior value** on a single item as a binary value to indicate whether his/her rating behavior is good or bad. When user u_j provides a rating to item I_i, if his/her rating is suspicious, the behavior value of u_j for item I_i, denoted by $Beh_{u_j}(i)$, is set to 0. Otherwise, $Beh_{u_j}(i)$ is set to 1. Assume that user u_j has rated $r + s$ items, where the behavior values for r items are 1 and for s items are 0, we define u_j's **combined behavior value** on these $r + s$ items as $Beh_{u_j}^{com} = \frac{r}{r+s+2}$, and the **behavior uncertainty** on these $r + s$ items as $Beh_{u_j}^{uncer} = \frac{2}{r+s+2}$.

Suppose that a user u_j has rated M items (i.e. $I_1, I_2, ..., I_i, ..., I_M$) in total, except item I_i, u_j has behavior value as 1 on r items and behavior value as 0 on s items. The trust value of user u_j on item I_i (i.e. $T_{u_j}(i)$), which indicates how much we could trust the rating provided by user u_j to item I_i, is calculated as

$$\begin{aligned} T_{u_j}(i) &= Beh_{u_j}^{com} * (1 - Beh_{u_j}^{uncer}) + Beh_{u_j}(i) * Beh_{u_j}^{uncer} \\ &= \frac{r}{r+s+2} * \frac{r}{r+s+2} + Beh_{u_j}(i) * \frac{2}{r+s+2} \end{aligned} \quad (4)$$

Finally, we detect the users with low trust values on items as malicious users. Instead of removing all the ratings provided by the malicious users, we only remove their ratings that yield low trust values. Specifically, for user u_j, if $T_{u_j}(i) < T_h$, user u_j's rating to item I_i is removed and u_j is marked as malicious user. Here, T_h is the trust threshold, which could be adjusted according to different application scenarios.

III. Experiment Results

To test the performance of the proposed scheme, we collected real user attack data against online rating systems through a cyber competition: CANT. The competition was launched on 05/12/2008 and lasted for 18 days. It successfully attracted more than 630 registered players and collected 826,980 valid submissions. In the competition, the normal rating data covered 300 products which were rated by 300 user IDs during 150 days. Players in the competition were required to submit attack strategies in which they can control up to 30 malicious user IDs to downgrade the reputation score of a particular product.

We group the attack submissions according to the number of malicious user IDs used in each submission (i.e. N_a), and then select 6 groups with N_a = 5, 10, 15, 20, 25, and 30, respectively. Since there are 300 normal users in the system, in these 6 groups of data, the malicious users have taken up 1.6%, 3.2%, 4.8%, 6.2%, 7.7% and 9% of the total user number. They are selected to represent attacks with very small, small, medium, large and very large number of malicious users. Attacks with a larger number of malicious users usually have stronger attack power and may cause larger attack impact.

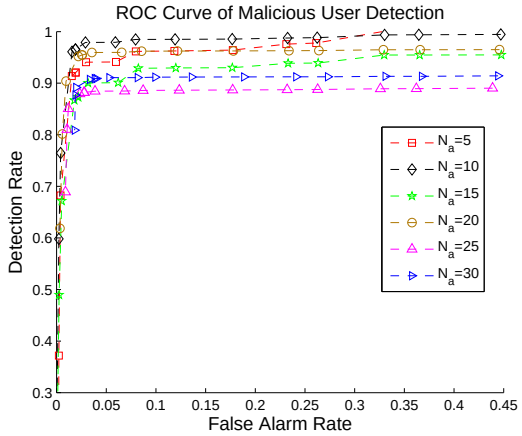

Fig. 1: Performance of malicious user detection for different malicious user number

Figure 1 shows the ROC curves for malicious user detection for different N_a values. The proposed scheme has demonstrated a consistent good performance in detecting malicious users. When the false alarm rate is around 5%, it yields high detection rate (i.e. $> 88\%$) for all attacks with different number of malicious users.

As a summary, the proposed scheme has achieved a very good performance in terms of accurately identifying malicious users. In the future work, we will continue to evaluate the performance of the proposed scheme in terms of recovering reputation score of the target items. Based on the accurate identification of malicious users, the proposed scheme has demonstrated a great potential to reduce the reputation distortion caused by malicious users' dishonest ratings, and keep the online reputation system a secure and fair marketplace.

References

[1] Yafei Yang, Qinyuan Feng, Yan Sun, and Yafei Dai, "Reputation trap: An powerful attack on reputation system of file sharing p2p environment," in *Proceedings of the 4th International Conference on Security and Privacy in Communication Networks*, Sep 2008.

[2] M. Abadi, M. Burrows, B. Lampson, and G. Plotkin, "A calculus for access control in distributed systems," *ACM Transactions on Programming Languages and Systems*, vol. 15, no. 4, pp. 706–734, 1993.

[3] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: defending against sybil attacks via social networks," in *Proceedings of the 2006 conference on Applications, technologies,architectures, and protocols for computer communications*, pp. 267-278, 2006.

[4] A. Jøsang and R. Ismail, "The beta reputation system," in *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

[5] J. Weng, C. Miao, and A. Goh, "An entropy-based approach to protecting rating systems from unfair testimonies," *IEICE TRANSACTIONS on Information and Systems*, vol. E89–D, no. 9, pp. 2502–2511, Sep 2006.

[6] A. Whitby, A. Jøsang, and J. Indulska, "Filtering out unfair ratings in bayesian reputation systems," in *Proceedings of the 7th Int. Workshop on Trust in Agent Societies*, 2004.

[7] J. Zhang and R. Cohen, "A personalized approach to address unfair ratings in multiagent reputation systems," in *Proc. of the Fifth Int. Joint Conf. on Autonomous Agents and Multiagent Systems (AAMAS) Workshop on Trust in Agent Societies*, 2006, pp. 89–98.

[8] P. Laureti, L. Moret, Y.-C. Zhang, and Y.-K. Yu, "Information filtering via iterative refinement," in *Europhysics Letters*, pp. 1006-1012, 2006.

[9] Y. Liu and Y. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Proc. of 2nd IEEE Int. Conference on Social Computing*, Aug 2010.

[10] G. Shafter, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.