Impact of Trust Management and Information Sharing to Adversarial Cost in Ranking Systems

Le-Hung Vu, Thanasis G. Papaioannou, and Karl Aberer

School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{lehung.vu,thanasis.papaioannou,karl.aberer}@epfl.ch

Abstract. Ranking systems such as those in product review sites and recommender systems usually use ratings to rank favorite items based on both their quality and popularity. Since higher ranked items are more likely selected and yield more revenues for their owners, providers of unpopular and low quality items have strong incentives to strategically manipulate their ranking. This paper analyzes the adversary cost for manipulating these rankings in a variety of scenarios. Particularly, we analyze and compare the adversarial cost to attack ranking systems that use various trust measures to detect and eliminate malicious ratings to systems that use no such a trust management mechanism. We provide theoretical results showing the relation between the capability of the trust mechanism in detecting malicious ratings and the minimal adversarial cost for successfully changing the ranking. Furthermore, we study the impact of sharing trust information between ranking systems to the adversarial cost. It is proved that sharing information between two ranking systems on common user identities and malicious behaviors detected can increase considerably the minimal adversarial cost to successfully attack the two systems under certain assumptions. The numerical evaluation of our results shows that the estimated adversary cost for manipulating the item ranking can be made significant when proper trust mechanisms are employed or combined.

Keywords: Trust; information sharing; open systems; adversarial cost; ranking systems; dishonesty detection.

1 Introduction

Ranking has become a popular and important feature of online business applications. A ranking system enables users to rate their favorite items based on item quality and also according to their own preferences. Items may represent services, products, sellable articles, digital content, or search results in different application scenarios. To facilitate the searching of users, these ratings are then used to rank a large number of items of the same category according to both their quality and popularity, e.g. ranking of digital content in social sites (Digg.com) or products in recommender systems (Amazon.com).

M. Nishigaki et al. (Eds.): IFIPTM 2010, IFIP AICT 321, pp. 108–124, 2010.

The impact of user online opinions on sales and profits is significant [1]. One can reasonably expect that items with higher ranks are more likely to be selected by clients and thus to produce more value for their providers. Consequently, there is a clear incentive for owners of unpopular and bad items to employ malicious identities to promote (i.e. "ballot-stuff") their own items and demote (i.e. "badmouth") competing ones to generate higher revenue. In real applications these issues are inevitable. For example, sellers can pay people for posting positive reviews on their products, as in [2] where Amazon reviews are bought with 65 cents each. Botnets can even be hired to conduct the attacks [3].

Regarding manipulation-resistance of ranking metrics, there have been a large number of works on studying resistance of Web page ranking algorithms, such as by throttling Web spams via link structure and link credibility analysis [4,5]. These works are applicable to large scale ranking systems that sort Web pages based on various criteria, such link quality and credibility of provider sites [6,7]. The application of trust mechanisms [8,9] to improve the robustness of a ranking system under adversarial attacks, such as ballot-stuffing and badmouthing is also well-explored [5,10]. However, the impact of the capability of a trust mechanism in detecting malicious ratings to the robustness of the ranking system using such mechanisms has not been analyzed yet.

To this end, we present in this paper an analytical approach to evaluate the robustness of a ranking system under attack by an intelligent adversary with limited resources. Particularly, we analyze the cost of an adversary to successfully manipulate the item ranking in smaller-scale systems, such as product review sites and recommender systems. The adversarial cost is estimated as the number of identities and ratings that need to be employed by the adversary to successfully change the ranks of specific targeted items. In practice, this cost may represent the cost of hiring people or botnets to post fake ratings on the targets [3]. We compare the adversary costs when specific trust mechanisms to eliminate biased ratings are employed or not. Thus, we provide theoretical results showing the relation between the capability of the trust mechanism being used to detect malicious ratings and the adversarial cost to attack a ranking system. By numerically evaluating our results, we show that the improvement in robustness of a ranking system using a trust mechanism with a given capability to detect dishonest ratings can be significant under certain assumptions.

Moreover, we extend our analysis to quantify the adversarial cost in an interesting scenario where two similar ranking systems share information regarding common users and the detection of malicious ratings. This scenario is realistic for the following reasons. On one hand, building applications that allows better exchanges of information on user activities with similar systems is an emerging trend adopted by many research and commercial initiatives, e.g., the on-going standardization of the OASIS committee on information exchange across reputation systems[1] and the OpenSocial API for better sharing information among

[1] `www.oasis-open.org/committees/orms`

online social networks. Commercial initiatives that are capable of collecting user activities across virtual communities are already available, some examples of which include Spokeo[2] and Reputation Defender[3]. On the other hand, malicious providers may want to publish their items in different systems for higher profits. To reduce cost, an adversarial provider may reuse a number of malicious identities across systems when posting bogus votes to manipulate the ranking of their items in different systems, e.g., by hiring only one botnet. Hence, by sharing the detection of malicious behaviors across systems, more malicious users are discovered and eliminated, which in turn helps to improve the robustness of the participating systems. We prove that, under certain realistic assumptions, two systems sharing information on common user identities and detected malicious ratings can increase the attack cost of an adversary considerably.

The remainder of this paper is organized as follows: in the next section, we describe the problem of ranking items in the presence of malicious raters. In Section 3, we analytically derive the minimum cost for the adversary to manipulate the ranking of the items under a trust mechanism that detects malicious votes with a certain effectiveness. In Section 4, we prove that the adversarial cost for manipulating the ranking of items increases when two systems exchange information regarding user identities and detected malicious ratings. Our results are numerically evaluated in Section 5, in Section 6 we discuss the related work before concluding the paper in Section 7.

2 Problem Formulation

Consider a ranking system with a set S of items (e.g. products or services), each having a binary static quality (good or bad). A user may rate the quality of the item after buying it. Let U be the set of honest raters. We denote as $r(u,s) \in \{1,0,-1\}$ the value of a rating from a user $u \in U$ for an item $s \in S$, where a value $r(u,s) = 0$ implies that u does not rate s. In general, a user $u \in U$ reports accurately the item quality. However, due to some observation noise, u may rate an item inaccurately with a small probability $0 < \varepsilon \ll 1$, e.g. a bad item is rated positively or vice versa. The items are ranked by their quality and popularity score (*QP-score*) $f(s)$ defined for any item $s \in S$ as:

$$f(s) = \sum_{u \in U} r(u,s)\,, \tag{1}$$

where a rating $r(u,s)$ is counted only once for each user u and each item s.

Let $S = \{s_i, 1 \le i \le M\}$ be the set of all items, where s_i has an original rank i according to the formula (1). Intuitively, $i < j$, or the item s_i is said to have a higher rank than s_j iff $f(s_i) > f(s_j)$.

The simple ranking function in (1) that only counts the number of positive and negative votes on an item is already effective to rank items in terms of their quality and popularity, provided that no adversary is present. In fact such a

[2] www.spokeo.com

[3] www.reputationdefender.com

metric has been used for ranking the digital contents on various Web 2.0 sites, e.g., to identify the most popular videos or blog entries. The use of sophisticated ranking metrics in more complex business applications, e.g. by considering credibility of the raters, belongs to the class of trust-based ranking functions that we will consider later on. We note that our approach to estimate the adversarial cost as presented in this work can even be extended to arbitrary ranking functions, although a closed-form solution cannot be easily obtained.

Suppose that there is an adversary who wants to boost the rank of an item s_k to the highest rank $k^* = 1 < k$. Herein, we use $k^* = 1$ to reduce the number of notations, but it is trivial to extend our analysis for any $k^* < k$. The same analytical reasoning could also be applied to the case that the adversary wanted to raise or lower the rank of a set of items instead of a single one. In order to promote item s_k, the adversary uses a set D of malicious user identities to post positive ratings on s_k and negative ratings on competing items, i.e. $s_i, 1 \leq i \leq k-1$. The total number of malicious ratings is C, and the cost of the adversary includes both components C and $|D|$.

For each item $s_i, 1 \leq i \leq k$, denote as U_i and D_i the set of honest and malicious users who rate on s_i, respectively. The number of ratings on an item s_i by a honest and malicious users are respectively $x_i = \mid U_i \mid$ and $y_i = |D_i|$. Depending the true quality (high or low) of s_i, the majority of x_i honest ratings on s_i would be positive or negative. Naturally, $\bigcup_{i=1}^{k} D_i = D$ and $\sum_{i=1}^{k} y_i = C$, since ratings items ranked lower than s_k does not help boosting the rank of the target item s_k but increase the cost of the adversary. Note that the adversary can observe, prior to his attack, the set U_i of any item s_i. We assume the worst case scenario where the adversary can estimate the numerical ranking score of every item, apart from the ranking, and thus can derive the cost C and $|D|$ to strategically change the ranking of items.

The system designer wants the ranking to reflect the true quality and popularity of items, so that the system is useful to its users. One naive approach that is often followed would be to simply ignore the presence of a possible adversary, and the items to be ranked according to the QP-score of each item s as in (1): $f_N(s) = \sum_{u \in U \cup D} r(u, s)$. To restrict the effect of the malicious ratings posted by the adversary, a preferable approach is to rank items based on the following trust-based QP score:

$$f_T(s) = \sum_{u \in U \cup D} r(u, s) t(u, s) \,, \tag{2}$$

where $0 \leq t(u, s) \leq 1$ is the estimated trustworthiness of the rating $r(u, s)$ and it is measured differently based on the trust management approach employed.

We focus in the comparison of the optimal cost of the adversary in terms of its minimal numbers of ratings C and malicious identities $|D|$, to successfully boost the rank of the item s_k in many situations where different QP scores $f_T(s), f_N(s)$ are used to rank items, and under different possible approaches to evaluate the trustworthiness of ratings. Note that without the adversary $D = \emptyset$, we have $f_T(s) = f_N(s) = f(s)$. Since $r(u, s)$ can be considered as a random variable, i.e., subject to observation noise or the honesty of the rating user, we estimate the

expected values $E[f(s)], E[f_N(s)], E[f_T(s)]$, whenever the exact rating $r(u,s)$ is unknown. Regarding the quality of the other items, we only consider the most important case where items in the competing set $s_i, 1 \leq i \leq k-1$ are of good quality (and thus they should be highly ranked for the benefit of the users). The other cases can be similarly analyzed.

3 Adversarial Cost under Trust-Based Ranking

3.1 Uniform Detection Capability of Malicious Ratings

Consider the system as described in Section 2, with approximate $x_i = |U_i|$ honest ratings (both positive and negative ones) on an item $s_i, i = 1, ..., |S|$. With the trust-based QP-score (2) as a ranking metric, the minimal cost of the adversary to manipulate the ranking is given by Proposition 1.

Proposition 1. *Suppose that the system uses a trust mechanism that can detect malicious ratings on any item with a probability* $0 < \gamma < 1$. *It is possible to design a ranking system in which the minimal adversarial cost, in expectation, to boost the rank of an item from* k *to* 1 *includes the cost of creating* $|D_T|$ *identities and posting* $C_T = |D_T|$ *ratings on the target item* s_k, *where:*

$$|D_T| = (x_1 + x_k)\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} \tag{3}$$

Proof. First, we prove that there exists a simple trust management approach that is capable of detecting malicious ratings on any item with a probability $0 < \gamma < 1$. The following naive trust management approach to define the trustworthiness $t(u,s)$ of a rating satisfies such a requirement (see Fig. 1):

- A trusted rater e is used to monitor the quality of a (uniformly) randomly selected set of items $E \subseteq S$, where $|E| = \gamma|S|$.
- For any $u \in U \cup D$, if there exists some item $s \in S$ such that the rating $r(u,s) \neq r(e,s)$, and $r(u,s)r(e,s) \neq 0$, we define $t(u,s) = 0$.
- Each remaining rating $r(u,s)$ has its trustworthiness proportional to the number of ratings with the same value. Formally, $t(u,s) = |Ut(s)|/|U(s)|$, where $U(s) \subseteq U \cup D$ is the group of users who rate on s, and $Ut(s) \subseteq U(s)$ is the users with ratings $r(u,s)$ on s.

Apparently, the above trust mechanism can detect malicious ratings on any item $s \in S$ with a probability γ, at the cost of the system designer evaluating $|E| = \gamma|S|$ items to learn of their true quality. Of course there may exist other trust mechanisms that are more cost-efficient, i.e., require the evaluation of less than $\gamma|S|$ services for a given capability of detection γ. The designing of such a trust mechanism is out of the scope of this analysis.

Recall that U_i and D_i are correspondingly the sets of honest and cheating raters on s_i. The trust-based QP score of an item $s_i, 1 \leq i \leq k$ is $f_T(s_i) = \sum_{u \in U_i \cup D_i} r(u,s_i)t(u,s_i)$. To effectively boost the rank of s_k, the adversary needs to post *at least* y_i negative ratings on each item $s_i, 1 \leq i \leq k-1$ and *at least* y_k positive ratings on the item s_k. The goal of the adversary is to ensure the

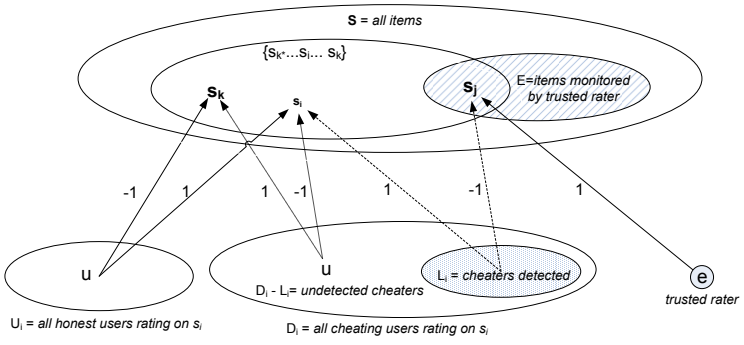

Fig. 1. Detection possibly malicious ratings on items by using a trusted rater

expected trust-based QP-score of the target item s_k to be as high as that of every other item of higher rank, i.e., $E[f_T(s_k)] \geq E[f_T(s_i)], 1 \leq i \leq k-1$.

Consider any item s_i, $1 \leq i \leq k-1$ with a good quality. Due to observation noise, among honest users U_i, a subset $U_i' \subseteq U_i$ may give unfair (negative) ratings on s_i. A smaller subset $U_i'' \subseteq U_i'$ may be detected by the trust management approach as cheater. Similarly, a subset of malicious users D_i who rate s_i negatively (to favor s_k) would be detected by the trust management mechanism. Denote as $L_i \subseteq D_i$ the set of malicious raters that are not detected. Then, users in the group $P_i = U_i - U_i'$ vote positively and those in the group $N_i = (U_i' - U_i'') \cup L_i$ vote negatively on s_i. Note that $P_i \cup N_i = (U_i - U_i'') \cup L_i$, as in Fig. 2(a). The trustworthiness $t(u, s_i)$ of a rating $r(u, s_i)$ is estimated as:

- For $u \in U_i'' \cup (D_i - L_i) : t(u, s_i) = 0$, i.e., users with erroneous observation and malicious users are marked as cheaters.
- For $u \in P_i = U_i - U_i'$: $t(u, s_i) = \frac{|P_i|}{|P_i \cup N_i|}$. Similarly, for $u \in N_i = (U_i' - U_i'') \cup L_i$, $t(u, s_i) = \frac{|N_i|}{|P_i \cup N_i|}$.

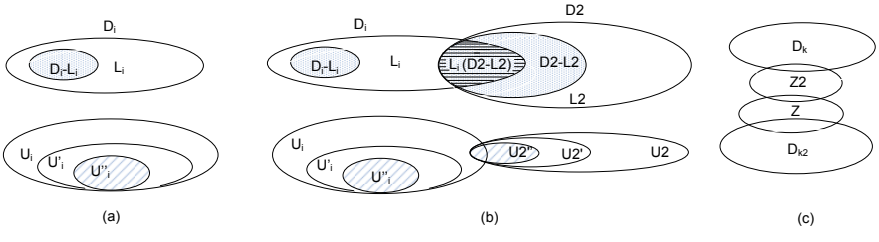

Fig. 2. (a) Venn diagram of the set of malicious and honest users detected by a trust mechanism. (b) The set of malicious and honest users detected by combining two trust management mechanisms. (c) Different sets of malicious users used by the adversary.

Eliminating ratings with 0 trustworthiness, i.e., those of users in the shaded parts of Fig. 2(a), the trust-based QP-score of any $s_i \in S, i = 1, ..., k-1$ becomes:

$$f_T(s_i) = \sum_{u \in P_i} 1.\frac{|P_i|}{|P_i| + |N_i|} + \sum_{u \in N_i} (-1)\frac{|N_i|}{|P_i| + |N_i|} = |P_i| - |N_i|$$

Since with a probability γ, malicious ratings on any item will be detected by the trust mechanism, we have:

- $E[|U_i'|] = |U_i|\varepsilon = x_i\varepsilon$, and $E[|U_i''|] = E(|U_i'|)\gamma = x_i\varepsilon\gamma$.
- $E[|D_i - L_i|] = E[\sum_{u \in D_i} 1_{\{u \text{ detected}\}}] = \sum_{u \in D_i} E[1_{\{u \text{ detected}\}}] = |D_i|\gamma$. It follows that $E|L_i| = |D_i|(1-\gamma) = y_i(1-\gamma)$.

As a result $E[|P_i|] = E[|U_i - U_i'|] = E[|U_i| - |U_i'|] = x_i(1-\varepsilon)$ and $E[|N_i|] = E[|(U_i' - U_i'') \cup L_i|] = E[|U_i'| - |U_i''| + |L_i|] = x_i\varepsilon - x_i\varepsilon\gamma + y_i(1-\gamma) = (1-\gamma)(x_i\varepsilon + y_i)$. Therefore, for any $1 \le i \le k-1$:

$$E[f_T(s_i)] = E[|P_i|] - E[|N_i|] = x_i(1-\varepsilon) - (1-\gamma)(x_i\varepsilon + y_i) = x_i(1 - 2\varepsilon + \varepsilon\gamma) - y_i(1-\gamma)$$

Similarly for the target item s_k, noting that honest users mostly rate negatively and malicious users rate positively on s_k, we have:

$$E[f_T(s_k)] = -E[|P_k|] + E[|N_k|] = -x_k(1-\varepsilon) + (1-\gamma)(x_k\varepsilon + y_k) = -x_k(1 - 2\varepsilon + \varepsilon\gamma) + y_k(1-\gamma)$$

The item s_k has a higher rank than s_i iff $E[f_T(s_k)] \ge E[f_T(s_i)]$, or:

$$y_k + y_i \ge (x_k + x_i)\frac{1 - 2\varepsilon + \varepsilon\gamma}{1-\gamma}$$

The minimal number of ratings the adversary needs to insert into the system is the solution of the following integer program:

$$\begin{aligned} C_T &= \min\{y_1 + y_2 + \ldots + y_k\} \\ \text{s.t. } & y_k + y_i \ge (x_i + x_k)(1 - 2\varepsilon + \varepsilon\gamma)/(1-\gamma), i = 1, ..., k-1 \end{aligned} \tag{4}$$

where all x_i, y_i are non-negative integers, x_is are fixed. One can also verify that as the first $k-1$ items are assumedly good, the number of ratings on them satisfies $x_i \ge x_{i+1}$, for $i = 1, ..., k-2$. This program has the following complete set of solutions[4]:

$$\begin{aligned} & y_k = (x_1 + x_k)(1 - 2\varepsilon + \varepsilon\gamma)/(1-\gamma) - d; y_1 = d; y_i = 0, 2 \le i \le k-1, \\ & \text{where } 0 \le d \le d_{max} = (x_1 - x_2)(1 - 2\varepsilon + \varepsilon\gamma)/(1-\gamma) \end{aligned} \tag{5}$$

Each solution above (for each $0 \le d \le d_{max}$) requires the adversary to post the same total number of ratings $C_T = \sum_{i=1}^{k} y_i = (x_1 + x_k)\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma}$. For each d, a corresponding attack strategy is to create at least $\max\{C_T - d, d\} = C_T - d$ identities[5]. Each of these $C_T - d$ identities posts a positive rating on the target item s_k; d identities are then reused to post negative ratings on the highest

[4] For clarity, we omit rounding operators $\lceil . \rceil$ from the right side of equations (5).

[5] Without loss of generality, we assume that $x_1 + x_k \ge 2(x_1 - x_2)$, hence $C_T - d_{max} \ge d_{max}$ and thus $\max\{C_T - d, d\} = C_T - d$.

ranked item s_1. With the attack strategy of $d = 0$, the adversary needs to create C_T identities, and the probability the attack is successful is $1-\gamma$. For $d > 0$, the adversary needs to create fewer $(C_T - d)$ identities, since he can use the same user to post ratings on both items s_1 and s_k. However, a strategy with $d > 0$ leads to higher chance that these identities are detected, and the probability that the attack is successful in this case becomes smaller, i.e., $(1-\gamma)^2 < 1-\gamma$. Formally, considering the expected gain and the risk of the adversary being detected, we can prove that the utility of the adversary is maximized at $d = 0$ in any of the two cases (1) γ is within a certain range or (2) the gain of the adversary if the attack is success is very large compared to its cost of creating d_{max} malicious identities. The proof is skipped due to space limitation. If we assume the case that the adversary cares most about the probability of success of the attack, the optimal strategy of the adversary is when $d = 0$, which incurs the following cost of creating at least $|D_T| = (x_1 + x_k)\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma}$ identities and posting at least $C_T = |D_T|$ ratings on the item s_k, as claimed by the proposition. □

In this paper, we refrain from presenting the analysis for the general case where an item $s_i, 1 \leq i \leq k$ has a true quality $q_i \in \{1, 0\}$ (high or low) for clarity reasons. This general result can be obtained by similar reasoning and by replacing the factor $x_1 + x_k$ in (3) with $(-1)^{1-q_1}x_1 - (-1)^{1-q_k}x_k$. Also by analogy, one can verify that the adversarial cost for promoting the target item to a desired rank $k^* < k$ can be obtained from (3) after replacing x_1 with x_{k^*}.

Following immediately from Proposition 1, we have an estimate of the extent of rank manipulation that can be done by an adversary.

Corollary 1. *If the system uses a trust mechanism that can detect malicious ratings on any item with probability $0 < \gamma < 1$, an adversary with capability to create at most $|D|$ identities and posts C ratings may manipulate the rank of a favorite item from the origin k to the highest rank $k^* \leq k$ defined by:*

$$k^* = \min_{k'=1}^{k}\{k' : (x_{k'} + x_k)\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma} \leq \min(C, |D|)\} \tag{6}$$

By similar reasoning, we obtain another result on the minimal adversarial cost when no trust mechanism is employed in the system (see Proposition 2).

Proposition 2. *In a system with no trust management mechanism to detect malicious users and eliminate their ratings, the minimal cost of the adversary to boost an item with rank k to rank 1 includes:*

- *The cost to create $|D| = (x_2 + x_k)(1 - 2\varepsilon)$ identities.*
- *The cost to post $C = (x_1 + x_k)(1 - 2\varepsilon)$ ratings on the two items s_1 and s_k.*

The optimal attack strategy is to post $d_{max} = (x_1 - x_2)(1 - 2\varepsilon)$ negative ratings on the top item s_1 and post $C - d_{max}$ positive ratings on the target item s_k.

The proof is similar to that of Proposition 1 for $\gamma = 0$. The difference is in the optimal attack strategy of the adversary. If the system uses no trust mechanism to detect malicious users and eliminate their ratings, the optimal strategy of the adversary to boost an item with rank k to rank 1 is attained at $d = d_{max}$, for

which the adversary needs to create only $C - d_{max}$ identities and uses them to vote negatively for s_1 and rate positively on the target item s_k.

From Proposition 2, we can also estimate to which extent an adversary with a fixed cost may manipulate the rank of his or her favorite items (Corollary 2).

Corollary 2. *Consider a system with no trust management mechanism to detect malicious users and eliminate their ratings. An adversary with capability to create at most $|D|$ identities and posts C ratings to the system may manipulate the rank of its favorite item from k to the highest rank $k^* \leq k$ defined by:*

$$k^* = \max\{\min_{k'=1}^{k}\{k' : (x_{k'+1} + x_k)(1-2\varepsilon) \leq |D|\}, \min_{k'=1}^{k}\{k' : (x_{k'} + x_k)(1-2\varepsilon) \leq C\}\} \quad (7)$$

Compared between the cost in Proposition 1 and Proposition 2, using a trust management mechanism that detects malicious ratings on any item with a probability γ would increase the minimal adversarial cost by some magnitudes:

$$|D_T|/|D| \simeq \frac{(x_1 + x_k)(1 - 2\varepsilon + \varepsilon\gamma)}{(x_2 + x_k)(1-2\varepsilon)(1-\gamma)} > 1 \quad (8)$$

$$C_T/C \simeq \frac{1 - 2\varepsilon + \varepsilon\gamma}{(1-2\varepsilon)(1-\gamma)} > 1 \quad (9)$$

Our analysis is general as the notion of γ include the capability of the trust mechanism to detect malicious on any item. There may exist other trust mechanisms that are more efficient in terms of guaranteeing a higher detection probability γ. These mechanisms might consider the reputation of the raters, credibilities of the item providers, and the correlation of ratings among raters to each others, etc. Designing such trust mechanism is, however, orthogonal to our work.

The cost of attacking the system also strongly depends on the set of votes by honest users, i.e., x_i. In systems where honest users outnumber the malicious users deployed by the adversary, manipulation of the trust-based ranking is much more costly to the adversary. Existing techniques to restrict the number of identities created by the adversary can be easily integrated to our analytical framework to restrict the capability of the adversary to manipulate the ranking.

3.2 Non-uniform Detection Capability of Malicious Ratings

Generally, the probability that the trust mechanism detects malicious ratings on different items may be non uniformed. For example, the trust mechanism may focus more on protecting of popular (and usually higher ranked) items, thereby increasing the probability of detecting unreliable ratings on these items. Let γ_i be the probability that malicious ratings on an item $s_i \in S$ are detected and eliminated. As a generalization of the analysis in Section 3.1, the optimal cost of the adversary to successfully manipulate the rank of the item s_k is the solution to the following integer program:

$$C_{ext} = \min\{y_1 + y_2 + \ldots + y_k\}$$

$$\text{s.t. } y_k(1-\gamma_k) + y_i(1-\gamma_i) \geq x_i(1 - 2\varepsilon + \varepsilon\gamma_i) + x_k(1 - 2\varepsilon + \varepsilon\gamma_k) \triangleq \phi_i, i = 1, ..., k-1$$

where all $0 < \gamma_i < 1$ are fixed, all x_i are fixed non-negative integers, and $x_i \geq x_{i+1}$, for $i = 1, ..., k-2$.

The probabilities γ_i are inherent to the trust mechanism, possibly determined by the system designer, while unknown to the adversary. The solution to the above optimization problem is the lower bound of the cost of the adversary. It is also our interest to evaluate which setting of $\gamma_1, ..., \gamma_k, ..., \gamma_{|S|}$ would result in a higher minimal cost of the adversary. Finding closed-form solutions for these cases is non-trivial and thus it is done numerically in Section 5.

4 The Benefits of Sharing Trust across Ranking Systems

This section presents the analysis of the adversarial cost in a system that uses an open trust management approach for detection and elimination of malicious ratings. That is, the system exchanges information on the identities of malicious users detected with another ranking system. Let $S2$ be the item set of the second system. Given any item $s'_j \in S2$, define $U2_j$ the set of honest users with ratings on s'_j, and $U2 = \bigcup_{s'_j \in S2} U2_j$. Also, let $D2_j$ be the set of malicious users with ratings on s'_j, and also define $D2 = \bigcup_{s'_j \in S2} D2_j$.

Assume that the second system uses another trust management approach that can detect malicious ratings on any item with a probability $0 < \gamma_2 < 1$. We assume that the two ranking systems are designed to automatically and reliably share the identities of malicious users detected to each other, and the system managers have low incentive to modify the software implementation to tamper such information. Fair and reliable information sharing between systems is an important issue that is beyond the scope of this paper and subject to future work. The identification of common users (in a privacy-preserving way) can be done via alias detection and entity resolution methods, e.g., based on credential attributes of the users. This problem is, however, orthogonal to the current analysis and thus is not further discussed.

For the case where two systems do not share any information, the adversary would need a set of D users to post a minimal number of C_T ratings to boost his favorites item s_k in the first system. Suppose that the goal of the adversary when attacking the second system is to boost the rank of an item $s'_{k_2} \in S2$ from k_2 to $k_2^* = 1$[6]. Then, the adversary would use another set of malicious users $D2$ to post a minimal number of C'_T ratings on his favorite items s_{k_2} in the second system. According to the analysis in Section 3.1:

$$C_T = (x_k + x_1)\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} =\mid D \mid \text{ and } C'_T = (x'_{k_2} + x'_1)\frac{1 - 2\varepsilon + \varepsilon\gamma_2}{1 - \gamma_2} =\mid D2 \mid \quad (10)$$

where $x'_i, i = 1, ..., k_2$ have similar meanings to those of the first system.

Suppose that the adversary is able to create up to $N =\mid D \cup D2 \mid$ identities in two systems for its malicious purposes. It is required that $N > \max\{C_T, C'_T\}$, otherwise with all N identities the adversary is still unable to attack both systems successfully. We will evaluate the benefit of sharing information between two

[6] Again we use $k_2^* = 1$ to reduce the notations without loss of generality of the analysis.

systems where such sharing is beneficial to both. That happens if the adversary does not have enough resources and needs to use a certain number of identities in both systems for its attacks, i.e., when $\max\{C_T, C'_T\} < N < C_T + C'_T$. Under this restriction, the adversary would use C_T among N identities to post C_T ratings on the first system. The posting of C'_T ratings in the second system will be done by employing: (1) the unused $N - C_T$ identities; (2) $C_T + C'_T - N$ among those C_T identities already used in the first system.

Hence, the cost of the adversary in case of no information sharing is:

- The cost of creating N identities, where $\max\{C_T, C'_T\} \leq N \leq C_T + C'_T$.
- The cost of posting $C_T + C'_T$ ratings in both systems.

When the two systems share trust evaluation results, the adversarial cost is:

- The same cost of N identities as in the case of not sharing information.
- The cost of posting $R_{\hat{T}}$ ratings, which would be defined later on.

We want to analyze how the adversarial cost in the case of sharing trust evaluation result differs from the case of not sharing any information, i.e., to quantify $R_{\hat{T}} - C_T - C'_T$.

Denote as $\tau_i = \mid U_i \cap U2 \mid, 1 \leq i \leq k$ the number of honest users who post ratings on s_i and also appear in the second system. We may approximate that $\tau_i = \mid U_i \cap U2 \mid \approx \tau / \mid S \mid, 1 \leq i \leq k$, where τ is the number of common honest users who post ratings in both systems. Similarly define $\tau'_i = \mid U2_i \cap U \mid \approx \tau / \mid S2 \mid, 1 \leq i \leq k_2$ the number of honest users who post ratings on $s'_i \in S2$ and also appear in the first system. The following main result gives us an estimation of the benefit of sharing information between the two systems.

Proposition 3. *Consider two ranking systems with capabilities γ, γ_2 of detection malicious ratings, where $0 < \gamma \leq \gamma_2 < 1$. Assume Δ be the number of identities the adversary needs to reuse in two systems, in the best case for the adversary, we have:*

$$0 \leq \Delta \leq \min\{(x_k + x_1)\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma}, (x'_{k_2} + x'_1)\frac{1 - 2\varepsilon + \varepsilon\gamma_2}{1 - \gamma_2}\} \tag{11}$$

If the two systems share trust evaluation information to each other, then the difference of the adversary cost to attack the two systems between two cases of sharing vs. non-sharing of information is bounded below by:

$$R_{\hat{T}} - C_T - C'_T > \frac{\Delta\gamma}{1 - \gamma} - \frac{\varepsilon\gamma_2(\tau_k + \tau_1)(1 - 2\varepsilon + \varepsilon\gamma)}{(1 - \gamma)^2} - \frac{\varepsilon\gamma(\tau'_{k_2} + \tau'_1)(1 - 2\varepsilon + \varepsilon\gamma_2)}{(1 - \gamma_2)^2} \tag{12}$$

Proof. We provide here a sketch of the proof (the full proof can be found in [11]). Let $z_i = \mid D_i \cap D2 \mid \leq y_i, 1 \leq i \leq k$ be the number of malicious raters who appear in both systems and rate an item $s_i \in S$ (of the first system). Similarly denote $z'_j = \mid D2_j \cap D \mid \leq y'_j, 1 \leq j \leq k_2$ the number of cheating users present in both systems and rate an item $s'_j \in S2$ (of the second system). Proceed as in Proposition 1, the minimal number of ratings $C_{\hat{T}}$ by the adversary to successfully attack the first systems is the solution to the following integer program:

$$C_{\widehat{T}} = \min\{y_1 + y_2 + ... + y_k\} \text{ subject to:}$$
$$y_k + y_i \geq (x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} + (z_k + z_i)\gamma_2, i = 1, ..., k-1$$
$$y_i \geq z_i, i = 1, ..., k$$

where $x_j, y_j, \tau_j, z_j, j = 1, ..., k$ are non-negative integers, all $x_i, \tau_i, z_i, i = 1, ...k$ is fixed, $x_i \leq x_j$, for $i \leq j, i, j = 1, ..., k-1$.

For $i = 1, ...k$, define $g_i = (x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma} - (z_k + z_i)(1 - \gamma_2)$. One may verify that any solution of the above program results in the same optimal number of ratings $C_{\widehat{T}} = \max\{0, \max_{i=1}^{k-1} g_i\} + \sum_{i=1}^{k} z_i$.

Similar to the previous section, if we assume that the adversary cares most about the probability of success of the attack, the optimal attack strategy is:

- for the item s_k: the adversary uses a set of users D_k from the first system and z_k identities from the second system to post $\widehat{y}_k =\mid D_k \mid + z_k$ ratings on s_k. We have $\mid D_k \mid = \max\{0, \max_{i=1}^{k-1} g_i\}$, and thus $C_{\widehat{T}} =\mid D_k \mid + \sum_{i=1}^{k} z_i$.
- for the items $s_i, i = 1, ..., k-1$: the adversary uses z_i identities from the second system to post z_i ratings on each s_i.

The set of malicious users to be used by the adversary in the first system is then $D = D_k \cup Z$, where $Z \subseteq D2$ is set of identities borrowed from the set of malicious users $D2$ in the second system. These borrowed identities are used by the adversary to post a total of $\sum_{i=1}^{k} z_i$ ratings on those items $s_i, i = 1, ..., k$. Likewise, the set of malicious users in the second system is $D2 = D_{k_2} \cup Z2$, where $Z2 \subseteq D$ is set of identities borrowed from the first system to rate on items in the second system. D_{k_2} is the set of malicious users who are only present in the second system and rate the target item $s'_{k_2} \in S2$.

The set of malicious users used by the adversary to attack both systems is thus $D_{\widehat{T}} = D_k \cup Z \cup D_{k_2} \cup Z2$. Fig. 2(c) illustrates the relation among different sets $D_k, Z, Z2, D_{k_2}$. Clearly, the malicious set $D_{\widehat{T}}$ is smallest iff $Z \subseteq D_{k_2}$ and $Z2 \subseteq D_k$. That is, the same malicious users in one system, e.g., D_{k_2}, are used to rate items in the other system, e.g., to rate item $s_i \in S, i = 1, ..., k$. Under such a situation, the total minimal number of identities the adversary needs to create in the two systems is $\mid D_{\widehat{T}} \mid =\mid D_k \mid + \mid D_{k_2} \mid$.

Similarly, the minimal number of ratings to be posted in the second system is $C'_{\widehat{T}} =\mid D_{k_2} \mid + \sum_{i=1}^{k_2} z'_i$. Thus the total cost of the adversary to attack both systems includes two cost: (1) to create $\mid D_{\widehat{T}} \mid$ identities and (2) to post $R_{\widehat{T}} = C_{\widehat{T}} + C'_{\widehat{T}}$ ratings in the two systems. Given a fixed number of identities N, the goal of the adversary is to determine the number of common users $z_i \geq 0, z'_j \geq 0, i = 1, ..., k, j = 1, ..., k_2$ such that $R_{\widehat{T}}$ is minimized. In other words:

$$R_{\widehat{T}} = \min\{\mid D_k \mid + \mid D_{k_2} \mid + \sum_{i=1}^{k} z_i + \sum_{j=1}^{k_2} z'_j\} \text{ subject to: } \mid D_{\widehat{T}} \mid =\mid D_k \mid + \mid D_{k_2} \mid = N$$
$$\text{where } \mid D_k \mid = \max\{0, \max_{i=1}^{k-1}\{(x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} - (z_k + z_i)(1 - \gamma_2)\}\}$$
$$\text{and } \mid D_{k_2} \mid = \max\{0, \max_{1 \leq i \leq k_2 - 1}\{(x'_{k_2} + x'_i - \varepsilon\gamma(\tau'_k + \tau'_i))\frac{1 - \varepsilon + 2\varepsilon\gamma_2}{1 - \gamma_2} - (z'_k + z'_i)(1 - \gamma)\}\}$$

Solving this program give us $R_{\widehat{T}} \geq -\frac{N\gamma}{1-\gamma} + \max_{i=1}^{k-1}\{\frac{f_i}{1-\gamma}\} + \max_{j=1}^{k_2-1}\{\frac{f'_j}{1-\gamma}\}$, where for simplicity we define $f_i \triangleq (x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma}, i = 1, ..., k-1$ and $f'_i \triangleq (x'_{k_2} + x'_i - \varepsilon\gamma(\tau'_{k_2} + \tau'_i))\frac{1-2\varepsilon+\varepsilon\gamma_2}{1-\gamma_2}, i = 1, ..., k_2 - 1$.

Since $\max\{C_T, C'_T\} \leq N < C_T + C'_T$, there are at least $\Delta = C_T + C'_T - N$ identities used by the adversary in the two systems, where:

$$0 \leq \Delta \leq \min\{C_T, C'_T\} = \min\{(x_k + x_1)\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma}, (x'_{k_2} + x'_1)\frac{1-2\varepsilon+\varepsilon\gamma_2}{1-\gamma_2}\}$$

The bound of Δ is for the best case of the adversary, when he can estimate the cost C_T, C'_T to successfully attack the two systems. Given Δ defined as above, with basic computations we obtain:

$$R_{\widehat{T}} - C_T - C'_T > \frac{\Delta\gamma}{1-\gamma} - \frac{\varepsilon\gamma_2(\tau_k + \tau_1)(1-2\varepsilon+\varepsilon\gamma)}{(1-\gamma)^2} - \frac{\varepsilon\gamma(\tau'_{k_2} + \tau'_1)(1-2\varepsilon+\varepsilon\gamma_2)}{(1-\gamma_2)^2} \quad (13)$$

and Proposition 3 follows naturally. □

The cost difference $R_{\widehat{T}} - C_T - C'_T$ in Proposition 3 mostly depends on the shortage of identities Δ of the adversary. The fewer number of identities the adversary has, the higher number of common identities it shall reuse across the two systems, and the more ratings it needs to insert into both systems to successfully manipulate the ranks of its favorite items. For most Δ and where the noise ε is negligible, it is apparent that $R_{\widehat{T}} - C_T + C'_T > 0$, or the adversarial cost to manipulate the ranking in both systems in the case of sharing trust information between the two systems is higher than the adversarial cost $C_T + C'_T$ where no information is shared. The capabilities of the two trust mechanisms in detecting malicious ratings, i.e. the probability γ, γ_2 also play an important rule in increasing this total adversarial cost. The sharing of information, however, may also lead to some false positives when estimating common users as cheating. This observation noise however plays a minor role, as the two negative terms on the right hand side of (12) are small, given small values of ε. Note that this cost difference $R_{\widehat{T}} - C_T - C'_T$ is estimated in the worst case where the adversary knows the common users (τ_i, τ'_i) and is aware of the effectiveness of the two system at detecting malicious activities (γ, γ_2) to develop an optimal strategy of placement malicious ratings in the two systems.

5 Numerical Evaluation

In this section, we numerically evaluate our results. All items including the target items are assumed to be good (but differ in popularity), which can be proven as even less costly for the adversary to promote them, and with the least difference between the number of ratings between items (hence the minimum adversarial cost is the lowest possible). The estimates are for $\varepsilon = 0.05$ and $M = |S| = 100$ items. There are $x_i = M - i$ honest ratings for each item with rank $1 \leq i \leq M$. Fig. 3 evaluates the increase in the minimal adversarial cost $|D_T|/|D|$ with respect to uniform detection capabilities γ of the trust management and with

various values of the original rank k and desired rank $k^* < k$ of the target item. We observe that even in this pessimistic scenario, the use of a trust mechanism with reasonable detection capability $\gamma = 0.5$ doubles the adversarial cost to manipulate the rankings in terms of the number of identities, irrespective of the original rank of the target item. The increase in adversarial cost by the number of malicious ratings C_T/C has a similar trend. Also, the raise of the adversarial cost for promoting the lowest ranked item can be achieved by increasing the detection capability of the trust mechanism being used γ (Fig. 3).

Next, we consider the impact to the minimal adversarial cost of a trust mechanism with non-uniform detection capabilities γ. For simplicity, we assume linear ascending and descending γ functions with respect to the item original rank and numerically solve the linear program of Section 3.2. The adversarial identities and ratings ratios ($|D_T|/|D|$ and C_T/C respectively) with respect to the initial item rank are depicted in Fig. 4. Thus, an ascending γ distribution increases the minimum adversarial cost for promoting lower ranked services. Then, we consider multiple *permutations* of the γ values for the items of various ranks and we observe that a trust mechanism that focuses more on detecting malicious ratings on lower ranked items increases the minimal adversarial cost to promote their ranking. In Fig. 5 the permutations corresponding to to the highest cost are those with γ ascending w.r.t the item rank.

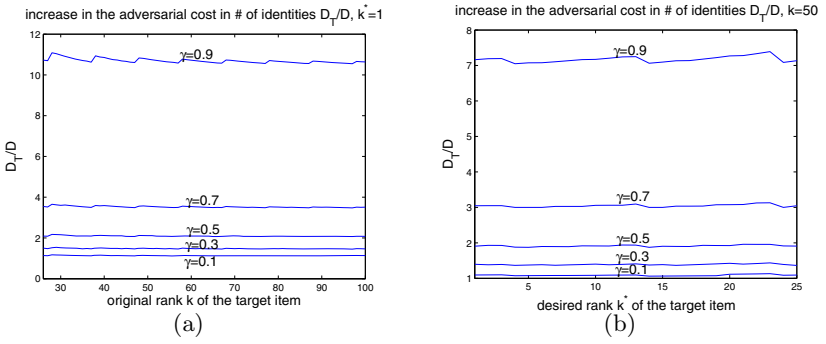

Fig. 3. Identities cost by a trust mechanism with different detection capabilities γ: (a) the target has variable original rank k and desired rank $k^* = 1$; and (b) the target has original rank $k = M/2$ and variable desired rank k^*

The impact of sharing trust information to the overall robustness of the two systems for an example case is given in Fig. 6, measured in the increase of adversarial cost (the number of ratings the adversary needs to insert into both systems). The two systems are assumed to use trust management mechanisms with similar detection capabilities $\gamma = \gamma_2$, have two similar item sets $|S| = |S2| = M$ with roughly $\tau = 10\%$ common honest users. The measurements are done in two representative cases where the target items have different original ranks in the two systems. The estimates are based on Eq. (12) in the worst case scenario with the least difference between the item popularity, $x_i = M - i, 1 \leq$

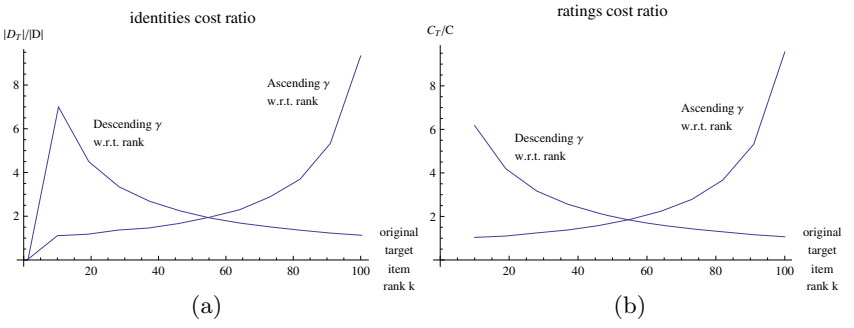

Fig. 4. Identities (a) and ratings (b) cost ratio for promoting an item with rank k with or without a trust mechanism employing an ascending or descending γ distribution

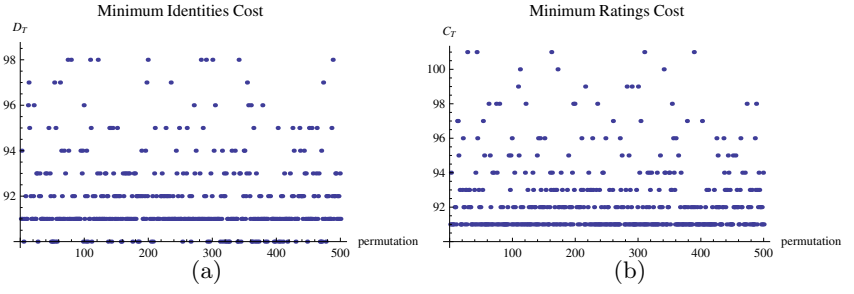

Fig. 5. Identities (a) and ratings (b) minimum cost for promoting an item initially ranked last with different γ distributions

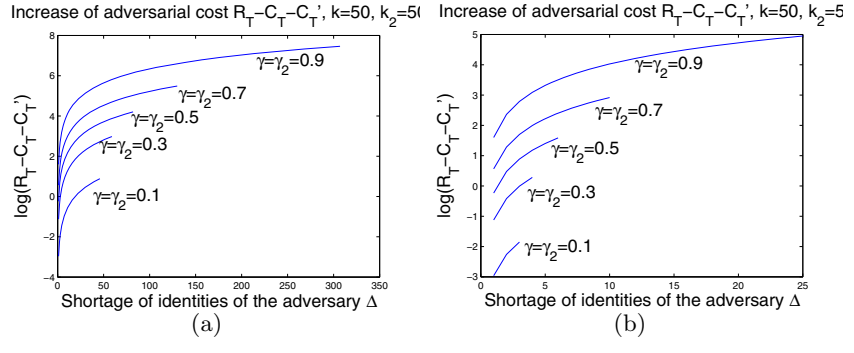

Fig. 6. Impact of the trust information sharing to the increase of adversarial cost (in log scale) where: (a) the (origin) rank of the target is average in both systems; (b) the rank of the target in one system is high. The results in other cases are similar.

$i \leq M, x'_j = M - j, 1 \leq j \leq M$. Observe that the sharing of information between two systems helps to significantly raise the total cost of the adversary to attack the two systems, thus strengthening both systems significantly. The conditions for this sharing of trust information to be beneficial to both systems, i.e., $log(R_{\widehat{T}} - C_T - C'_T) > 0$ are: (1) the detection capabilities of the two systems are sufficiently high, and (2) the resources of the adversary are limited, e.g., $\gamma, \gamma_2 > 0.5$ and $\Delta > 5$ in the case of Fig. 6.

6 Related Work

The works most related to ours include existing research on resilience of Web page ranking algorithms against Web spams, via link structure and credibility analysis, namely [4, 5]. The use of trust and reputation mechanisms to minimize the influence of adversarial attacks in ranking systems has also attracted much effort [8, 9]. EigenTrust [12] presents a global trust metric to measure the credibility to a node in a network based on inter-connecting links among nodes. Other works, as [5], use reputation-based trust management techniques to improve the robustness of ranking systems but with little analysis on the impact of trust mechanisms to the adversarial cost for strategic manipulation the system. A more recent work [10] studies vulnerabilities and attacks by an adversary with a given cost to voting systems and propose defense mechanisms based on item popularity. This work is different from ours as it only considers the binary voting result on item quality while our work is more general: we consider ranking systems that use both popularity and quality of items as ranking metrics.

7 Conclusion

This paper analyzes the minimum adversarial cost to manipulate the ranking of items in systems where a trust mechanism is employed for detecting unfair and biased ratings. We provide theoretical results showing the relation between the capability of the trust mechanism being used to detect malicious ratings and the minimal adversarial cost to successfully attack a ranking system. Moreover, we have proved that, under certain realistic assumptions, two systems with shared information on malicious user activities can increase the minimal successful attack cost of an adversary considerably. Our analysis indicates that the cost of the system designer to prevent attacks from the adversary with a certain power, is related to the cost of implementing a trust management mechanism with a certain capability of detecting malicious rating behaviors. The analytical framework in the paper can be extended to estimate the robustness of more complex ranking score metrics against the adversary. It may also be our interest to analyze the cost and the influence on the final ranking result in presence of many competing adversaries with different powers.

References

1. Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research 43(9) (2006)
2. Parsa, A.: Belkins Development Rep is Hiring People to Write Fake Positive Amazon Reviews (2009)
3. Namestnikov, Y.: The economics of Botnets (2009), http://www.viruslist.com/analysis?pubid=204792068
4. Caverlee, J., Webb, S., Liu, L., Rouse, W.B.: A parameterized approach to spam-resilient link analysis of the web. IEEE Trans. Parallel Distrib. Syst. 20(10), 1422–1438 (2009)
5. Gyongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), pp. 271–279 (2004)
6. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University (1998)
7. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (1999)
8. Golbeck, J.: Trust on the world wide web: A survey. Foundations and Trends in Web Science 1(2), 131–197 (2006)
9. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decis. Support Syst. 43(2), 618–644 (2007)
10. Feng, Q., Sun, Y., Liu, L., Yang, Y., Dai, Y.: Voting Systems with Trust Mechanisms in Cyberspace: Vulnerabilities and Defenses. IEEE Transactions on Knowledge and Data Engineering (to appear, 2010)
11. Vu, L.H., Papaioannou, T.G., Aberer, K.: Impacts of trust management and information sharing to adversarial cost in ranking systems. Technical Report LSIR-REPORT-2010-001 (2010), http://infoscience.epfl.ch/record/143071
12. Kamvar, S.D., Schlosser, M.T., Molina, H.G.: The EigenTrust algorithm for reputation management in P2P networks. In: Proc. of WWW 2003 (2003)