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 comparison 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 bad items have strong incentives to manipulate the ranking in favor of their own items. 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 trust mechanism. We provide theoretical results showing the relation between the capability of the trust mechanism in detecting malicious ratings and the minimum adversary 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 significantly increase the adversarial cost to attack any of them under certain assumptions. Our results are numerically evaluated showing that the estimated adversary cost for manipulating the item ranking can be made significant when proper trust mechanisms are employed or combined.

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), of products in recommender systems (Amazon.com) and in search engines (www.google.com/products).

The impact of user online opinions on sales and profits is well-known to be 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.However, 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] for Amazon reviews for 65 cents each, or even hire botnets to conduct the attack.

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 spam via link structure and link credibility analysis [3, 4]. 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 [5, 6]. The application of trust mechanisms [7, 8] to improve the robustness of a ranking system under adversarial attacks, such as ballot-stuffing and badmouthing is also well-explored [4, 9]. 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, in this paper, we present 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 rating sites and recommendation 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. We compare this cost when specific trust mechanisms to eliminate biased ratings are employed or not. The improvement in robustness of a ranking system using a trust mechanism with a given capability to detect dishonest ratings is numerically evaluated to be significant under certain assumption.

Moreover, we extend our analysis in a more interesting scenario where two similar ranking systems share information regarding user identities and detect malicious ratings mutually. This scenario is realistic for two reasons. First, collecting and exchanging information regarding identities and activities of users across systems are feasible in practice. Commercial initiatives for aggregating online reputation information related to a person across different sites have become increasingly popular, such as Online Reputation Monitor (reputation.distilled.co.uk), Reputation Manager (www.reputationmanager.com), or Reputation Defender (www.reputationdefender.com). The OASIS committee has also proposed standards for information exchange and reuse across reputation systems (www.oasis-open.org/committees/orms). This standardization effort, if realized, would further facilitate the integration of similar reputation-aware systems. Second, malicious providers may want to publish their items in similar systems for maximizing profits, while having limited resources. This limitation requires that the adversary reuses a number of malicious identities across systems when posting ratings to manipulate the ranking of its targeted items. Hence by sharing information on the detection of malicious behaviors between two similar systems, more ma-

licious users are discovered and eliminated, which in turn helps to improve the robustness of both systems.

The contribution of this paper is twofold. First, 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. We also provide numerical evaluation of this relation in various settings. Second, we extend the analysis to quantify the adversarial cost of attacking two similar ranking systems sharing information on user identities and the detection of malicious ratings. We show that, under certain realistic assumptions, two systems with shared information regarding common user identities and trust evaluation result can significantly increase the attack cost of an adversary. The analytical framework in the paper can also be extended to estimate the robustness of more complex ranking metrics under the presence of an adversary. 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 ratings. In Section 3, we analytically derive the minimum cost for the adversary to manipulate the ranking of the items, when a trust mechanism that detects malicious votes with a certain effectiveness is employed. In Section 4, we analytically prove that the adversarial cost for manipulating the ranking of items increases when two systems exchange information regarding user identities and detect malicious ratings. Our results are numerically evaluated in Section 5, while in Section 6 we discuss the related work. Finally, in Section 7, we conclude our work.

2 Problem Formulation

Consider a ranking system with a set of items S, each has a binary static quality (good/bad). One item can be a representation (description) of similar articles/services provided by a seller and thus can be sold for different users. Let U be the set of all rating users who are honest. Denote as $r(u, s) \in \{1, 0, -1\}$ the value of a rating by any user $u \in U$ on an item $s \in S$, where a value $r(u, s) = 0$ implies that u does not rate s. A user $u \in U$ in general reports honestly on the item quality. Due to some observation noise, with a small probability $0 < \varepsilon \ll 1$, u may rate an item incorrectly, i.e., a bad item is sometimes rated positively and a good item may be rated negatively. Items are then 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 $S = \{s_i, 1 \leq i \leq M\}$ be the set of all items where s_i has an original rank i according to the metric (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 metric $f(.)$ counting number of positive and negative votes on an item is usually used in existing systems to rank items in term of their quality and popularity. The use of more sophisticated metrics considering timestamp, and credibility of raters corresponding to other trust-based ranking metrics that will be studied later on.

Now consider an adversary who wants to boost the rank of an item s_k to the highest rank $k^* = 1 < k$. Herein we use $k^* = 1$ only to reduce the number of notations, it is trivial to extend the analysis to any $k^* < k$. It is also straightforward to use our analytical framework to the case where the adversary wants to boost or lower the rank of a set of items instead of a single one. In order to do so, the adversary uses a set of malicious users D to post positive ratings on s_k (the boosted item) and negative ratings on those items $s_i, 1 \leq i \leq k-1$ (the competing set). The total number of malicious ratings is C, and the cost of the adversary includes both components C and $\mid D \mid$.

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 $y_i = \mid D_i \mid$. Depending the true quality (good or bad) 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. We assume that the adversary knows the honest user set U_i of any item s_i and can estimate the actual rank of every item. The system manager, however, does not know the sets U_i, D_i, and the target item s_k.

The system designer wants the ranking to reflect the true quality and popularity of items as observed by the honest users, so that new users do not have bad experience in using the system to choose their items. One naive approach is to simply ignore the presence of a possible adversary, and items are 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 malicious ratings posted by the adversary, a better way is to rank items based on a trustworthiness measure of each rating, namely for each item s, the following trust-based QP score of is used instead of (1):

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

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

The goal of this work is to compare the optimal cost of the adversary, in terms of its minimal numbers of ratings C and malicious identities $\mid D \mid$, 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 given different possible approaches to evaluate the rating trustworthiness. Note that in absence of the adversary $D = \emptyset$, we have $f_T(s) = f_N(s) = f(s)$, thus the trust-based and naive quality score are in fact generalization of the normal quality score $f(s)$. Since $r(u, s)$ can be considered as a random variable, i.e., subject to observation noise or behavior of the rating user, we consider the expected value $E[f(s)], E[f_N(s)], E[f_T(s)]$, whenever the exact rating $r(u, s)$ is unknown.

We only consider the most important cases where items in the competing set $s_i, 1 \leq i \leq k-1$ are of good quality (and thus shall be ranked highly for the benefits of the users). As point out in the analysis, the other cases are similar and thus skipped for space limitation.

3 Adversarial Cost to Influence the Trust-based Ranking

Consider the system in Section 2, with approximate $x_i =\mid U_i \mid$ honest (both positive and negative) ratings on an item $s_i, i = 1, ..., \mid S \mid$. With the trust-based QP-score (2) as a ranking metric, Proposition 1 gives the minimal adversarial cost to manipulate the ranking.

Proposition 1. *Suppose the system uses a trust mechanism that detects malicious ratings on any item with 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* $\mid D_T \mid$ *identities and posting* $C_T =\mid D_T \mid$ *ratings on the target item* s_k, *where:*

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

Proof. First, we prove that there exists a simple trust management approach that detects 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 randomly selected set of items $E \subseteq S$, where $\mid E \mid= \gamma \mid S \mid<\mid S \mid$.
- for any $u \in U \cup D$, if there exists some item $s \in S$ such that the ratings $r(u, s)r(e, s) \neq 0, r(u, s) \neq r(e, s)$, we define $t(u, s) = 0$.
- for each of the remaining ratings $r(u, s)$, the trustworthiness is proportional to the number of ratings with the same value. Formally, $t(u, s) =\mid Ut(s) \mid / \mid U(s) \mid$, 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.

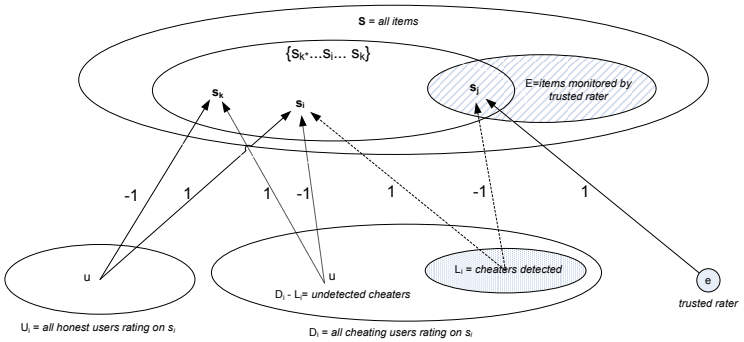


Fig. 1. Detection unfair ratings on items by using a trusted rater.

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 $\mid E \mid= \gamma \mid S \mid$ 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 $\gamma \mid S \mid$ services for a given capability of detection γ, but the designing of such trust mechanism is not the focus of this work.

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 expected trust-based QP-score of the target item s_k higher than that of every 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. 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 (based on erroneous ratings). Among those malicious users D_i who rate s_i negatively (to favor s_k), a subset of them 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 $u \in N_i = (U_i' - U_i'') \cup L_i$, $t(u, s_i) = \frac{|N_i|}{|P_i \cup N_i|}$.

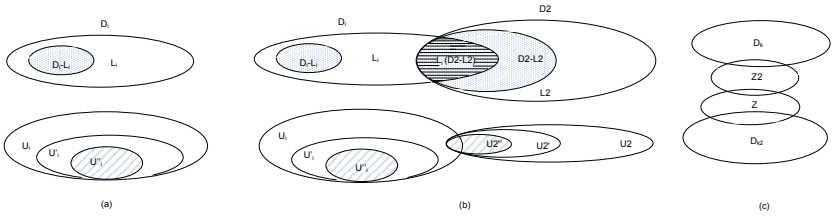


Fig. 2. (a) Venn diagram of the set of malicious and honest users detected by using a trust mechanism; (b) The set of malicious and honest users detected by combining two trust management systems; (c) Different sets of malicious users used by the adversary to attack the two systems

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:

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

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

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

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

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

Similarly, for the target item s_k, noting honest users mostly rate s_k negatively while malicious users rate it positively, we have:

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

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

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

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

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

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 $x_i \geq x_{i+1}$, for $i = 1, ..., k-2$. This program has the following complete set of solutions[1]:

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

Each solution above (for each $0 \leq d \leq d_{max}$) 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[2]. Each of these $C_T - d$ identities posts a positive rating on the boosted item s_k. The adversary also uses d identities to post negative ratings on the highest 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

[1] For presentation clarity, we skip the rounding operators $\lceil . \rceil$ or $\lfloor . \rfloor$ when mentioning quantities involving integer values, e.g., the number of ratings and identities.

[2] We assume without loss of generality that $x_1 + x_k \geq 2(x_1 - x_2)$, hence $C_T - d_{max} \geq d_{max}$ and thus $\max\{C_T - d, d\} = C_T - d$

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 $\mid D_T \mid = (x_1 + x_k)\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma}$ identities and posting at least $C_T = \mid D_T \mid$ ratings on the item s_k, as claimed by the proposition. □

For simplicity of the explanation, in this paper we do not present the analysis for the general case where an item $s_i, 1 \leq i \leq k$ has a true quality $q_i \in \{1, 0\}$ (good or bad). By similar reasoning, this general result can be obtained 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 by replacing x_1 in (3) with x_{k^*} to obtain the increase in the adversarial cost to boost the rank of the target item to any desired rank $k^* < k$.

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 $\mid D \mid$ 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:*

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

By similarly reasoning, we have another result on the minimal adversarial cost to manipulate ranks in a system without any trust management mechanism to eliminate malicious ratings (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 $\mid D \mid = (x_2 + x_k)(1 - 2\varepsilon)$ 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 $\mid D \mid$ 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:*

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

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:

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

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

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, an orthogonal issue to our analysis. For example, one way to increase the detection probability γ is to use trust-distrust propagation as in our previous work [10], where a rater is evaluated as cheating if some of its ratings are evaluated to be incorrect and/or similar to the ratings of those users already discovered as malicious.

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 in can be easily integrated to our analytical framework to restrict the capability of the adversary to manipulate the ranking.

Using a Trust Management Mechanism with Non-uniform Capability of Detection 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, 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 is done numerically in Section 5. The baseline for this evaluation is the adversary cost of a selection of items with equal probability for monitoring $\gamma_i = \gamma$ is used (3,**??**).

Let $i_0 = argmax_{1 \leq i \leq k-1} \phi_i$. We can find a closed-form solution for the above optimization program for the case $\gamma_{i_0} \geq \gamma_k$. This case corresponds to, for example, when the system designer focuses more on protection of higher ranked items rather than on the lower rank ones that likely includes the target item s_k. In this case the optimal strategy of the adversary is to create at least $\mid D_{ext} \mid = \frac{\phi_{i_0}}{1-\gamma_k}$ identities, and use these identities to post $C_{ext} = \mid D_{ext} \mid = \frac{\phi_{i_0}}{1-\gamma_k}$ ratings on the item s_k (see Appendix A for a detailed analysis). In a simple case where $\gamma_i \geq \gamma_{i+1}, 1 \leq i \leq \mid S \mid -1$, it follows that $\phi_{i_0} = \phi_1 = (x_1 + x_k)(1 - 2\varepsilon + \varepsilon\gamma_1)$. Compared to the cost in Proposition 1, the current system would increase the adversarial cost by a magnitude:

$$\frac{C_{ext}}{C_T} = \frac{\mid D_{ext} \mid}{\mid D_T \mid} \simeq \frac{\phi_{i_0}}{1-\gamma_k} . \frac{(1-\gamma)}{(x_1 + x_k)(1 - 2\varepsilon + \varepsilon\gamma)} = \frac{(1-\gamma)(1 - 2\varepsilon + \varepsilon\gamma_1)}{(1-\gamma_k)(1 - 2\varepsilon + \varepsilon\gamma)}. \quad (10)$$

Therefore, the system that can better protect higher ranked items may be either stronger or weaker than another system that can detect malicious ratings on any item with the same probability γ, depending on γ_1, γ.

In a general case, we need to quantify the adversarial cost of manipulate the ranking with different settings $\gamma_i, 1 \leq i \leq \mid S \mid$ to evaluate whether it is better to use a trust mechanism to protect higher ranked items or not, i.e., the detection probability γ_i is higher for small i. Finding closed-form solutions for these cases is non-trivial and thus 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 regarding 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 $0 < \gamma_2 < 1$. In this paper we assume that the two ranking systems are designed to automatically and reliably share the identities of malicious users detected to each other, and system managers have little incentives to modify the software implementation to tamper such information. Fair and reliable information sharing between systems are an important issue yet beyond the scope of this paper.

For the case where two systems do not share any information, the adversary would need a set of D users to post 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 $k_2^* = 1$[3]. Then, the adversary would use another set of malicious users $D2$ to post C'_T ratings on his favorite items s_{k_2} in the second system. According to the analysis in section 3:

$$C_T = (x_k + x_1)\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} =\mid D \mid \text{ and } C'_T = (x'_{k_2} + x'_1)\frac{1 - 2\varepsilon + \varepsilon\gamma_2}{1 - \gamma_2} =\mid D2 \mid (11)$$

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 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 $N < C_T + C'_T$. 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 $R_{\widehat{T}}$ ratings, which would be defined later on.

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

Denote $\tau_i =\mid U_i \cap U2 \mid, 1 \leq i \leq k$ as 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, where:*

$$0 \leq \Delta \leq \min\{(x_k + x_1)\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma}, (x'_{k_2} + x'_1)\frac{1 - 2\varepsilon + \varepsilon\gamma_2}{1 - \gamma_2}\} \quad (12)$$

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

[3] Again, $k_2^* = 1$ reduces the notations without loss of generality of the analysis.

sharing vs. non-sharing of information is bounded below by:

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

The proof is given in Appendix B of the paper.

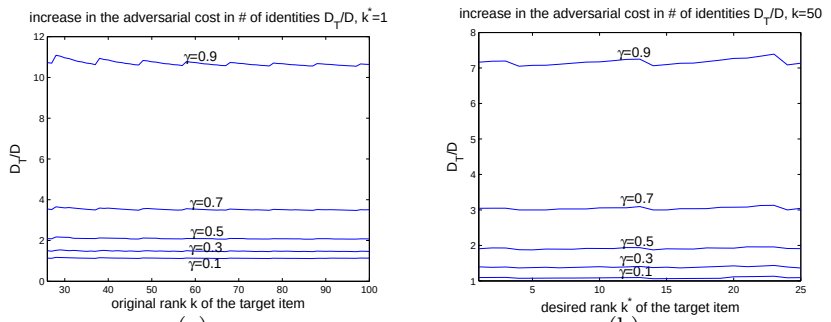
Proposition 3 gives us an estimate of the difference $R_{\widehat{T}} - C_T - C'_T$. This conclusion is hold under the extremely worst case assumptions: the adversary knows other common users (τ_i, τ'_i), knows of which system is better at detecting malicious activities (γ, γ_2) to develop an optimal strategy of ratings and placement of malicious entities in the two systems.

The increase in the cost $R_{\widehat{T}} - C_T - C'_T$ 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. These observation noise however plays a minor role, as the two negative terms on the right hand side of (68) are small, given small values of ε.

5 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).

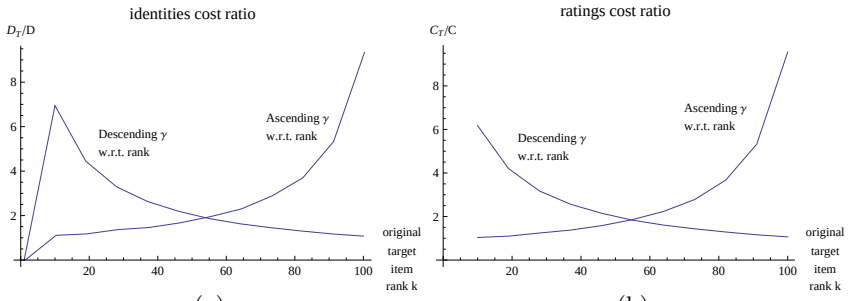
Next, we consider the impact to the minimal adversarial cost of a trust mechanism with non-uniform detection capabilities γ. For simplicity, we assume linear ascending and descending γ functions with respect to the item original rank and



(a) (b)

Fig. 3. Impacts of a trust mechanism with different detection capabilities γ to the adversarial cost measured in number of malicious identities to be created: (a) the target has various original rank k, the desired rank $k^* = 1$; and (b) the target has original rank $k = M/2$ and various desired ranks k^*.

numerically solve the linear program (14). The adversarial identities and ratings ratios ($\mid D_T \mid / \mid D \mid$ and $\mid C_T \mid / \mid C \mid$ 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. Also, considering multiple different γ distributions (Fig. 5), we observe that a trust mechanism that focuses more on protection more of lower ranked items increases the minimal adversarial cost to promote ranking of these items.



(a) (b)

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

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 to similar item sets $\mid S \mid = \mid S2 \mid = M$ with roughly $\tau = 10\%$ common honest users. The measurements are done in three representative cases where the target items have different original ranks in the two systems. The estimates are based on Eq. (68) in the worst case scenario where there is the least difference between the popularity of the items,

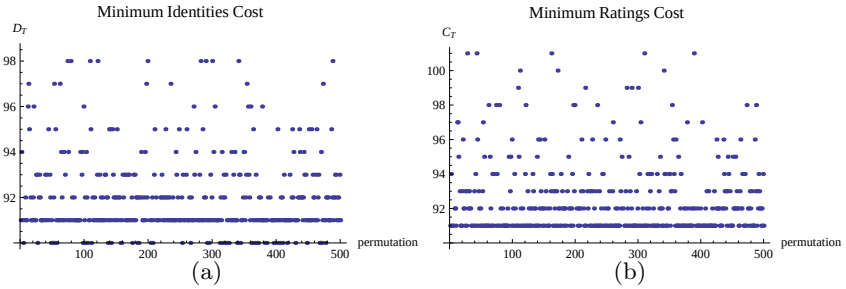


Fig. 5. Identities (a) and ratings (b) minimum cost for promoting a service initially ranked last with different γ distributions.

$x_i = M - i, 1 \leq i \leq M, x'_j = M - j, 1 \leq j \leq M$. 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.

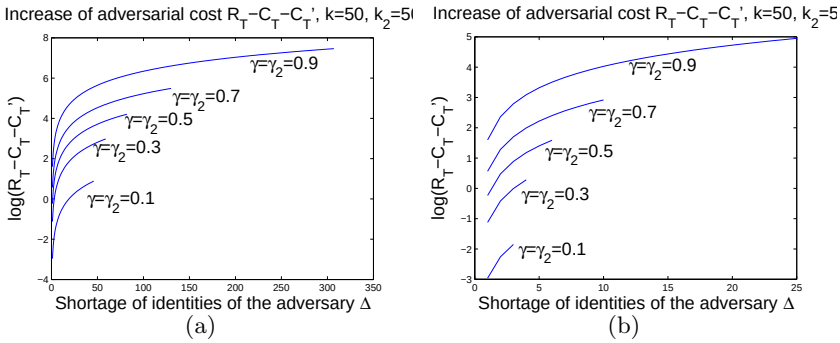


Fig. 6. Impact of the trust information sharing to the increase of adversarial cost (in logscale) where: (a) the (origin) rank of the target is average in both systems; (b) the rank of the target in one system is high. The results in other cases are similar.

6 Related work

The most related work to this paper is those study the resistance of Web page ranking algorithms and prevention of Web spams, e.g., via link structure and link credibility analysis [3,4]. The use of trust and reputation mechanism to minimize the influence of adversarial attacks in ranking systems has also attracted much effort [7,8]. EigenTrust [11] presents a global trust metric to measure the credibility to a node in a network based on inter-connecting links among nodes. This trust mechanism can be used to build an attack-resistant ranking system similar to the PageRank approach [5] to ranking Web pages. Other works suggested using reputation-based trust management technique to improve robustness of

ranking systems include [4, 10]. However, there seems to be little analysis on impacts of trust mechanisms to the cost of manipulation a ranking system. The most related work in this aspect is [9], where the authors study 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 since it only considers the binary voting result on item quality. Our work is more general: we consider ranking systems that use both popularity and quality of items as ranking metrics. We also analyze and quantify the cost of targeted adversarial attacks to manipulate the rankings in different scenarios, where the systems use trust management mechanisms with different detection capabilities.

7 Conclusion

This paper analyzes the cost of manipulation of items ranking in systems with different capabilities of detecting unfair and biased ratings. We provide theoretical results showing the role of the capability of the trust mechanism being used to the cost for the adversary to successfully attack manipulate the ranking, and numerically evaluate this relation in various settings. Furthermore, we analyze and verify numerically that two ranking systems with shared information regarding common user identities and detection of malicious behaviors may help to increase the attack cost of an adversary. This claim holds mainly because creating identities is costly and the adversary may need to reuse a number of malicious users across two systems to save the total cost. According to our analysis, under certain assumptions sharing information among systems would circumvent common cheating users, making it harder for an adversary to inverse the ranking produced by the participating systems, and strengthening them significantly.

A Adversarial cost for systems using trust mechanism with non-uniform probability of malicious detection

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, the optimal cost of the adversary to successfully manipulate the rank of the item s_k is the solution to the following integer program:

$$\begin{aligned} &C_{ext} = \min\{y_1 + y_2 + \ldots + y_k\} \\ &\text{s.t. } y_k(1-\gamma_k) + y_i(1-\gamma_i) \geq x_i(1-2\varepsilon+\varepsilon\gamma_i) + x_k(1-2\varepsilon+\varepsilon\gamma_k) \triangleq \phi_i, i = 1, ..., k-1 \quad (14) \end{aligned}$$

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

Define $i_0 = argmax_{1 \leq i \leq k-1} \phi_i$, we have:

$$y_k(1-\gamma_k) + y_{i_0}(1-\gamma_{i_0}) \geq \phi_{i_0} = \max_{1 \leq i \leq k-1} \phi_i \quad (15)$$

We can easily find a closed-form solution for the case where $\gamma_{i_0} \geq \gamma_k$. This case corresponds to, for example, when the system designer focuses more on

protection of higher ranked items rather than on the lower rank ones that likely includes the target item s_k. For the other case where $\gamma_{i_0} < \gamma_k$, finding the solution set is non-trivial and thus should be done numerically.

Given that $\gamma_{i_0} \geq \gamma_k$, let $\widehat{y}_i, i = 1, ..., k$ is a solution to the above integer program, it follows from (15) that:

$$(\widehat{y}_k + \widehat{y}_{i_0})(1 - \gamma_k) \geq \widehat{y}_k(1 - \gamma_k) + \widehat{y}_{i_0}(1 - \gamma_{i_0}) \geq \phi_{i_0} \tag{16}$$

$$\Rightarrow \widehat{y}_k + \widehat{y}_{i_0} \geq \frac{\phi_{i_0}}{1 - \gamma_k} \tag{17}$$

Apparently, $\widehat{y}_k = \frac{\phi_{i_0}}{1-\gamma_k}, \widehat{y}_i = 0, i \neq k$ is a solution. Therefore, the complete set of solutions is any $\widehat{y}_i, i = 1, ..., k$ such that:

$$\widehat{y}_k = \frac{\phi_{i_0}}{1 - \gamma_k} - d \tag{18}$$

$$0 \leq \widehat{y}_{i_0} = d \leq d_{max}, \text{ other } \widehat{y}_i = 0 \tag{19}$$

Each solution results in the same optimal number of ratings $C_{ext} = \sum_{i=1}^{k} \widehat{y}_i = \frac{\phi_{i_0}}{1-\gamma_k}$. The value d_{max} can be found by the following constraint:

$$\widehat{y}_k(1 - \gamma_k) + \widehat{y}_i(1 - \gamma_i) \geq \phi_i, i = 1, ..., k-1, i \neq i_0 \tag{20}$$

$$\Rightarrow (\frac{\phi_{i_0}}{1 - \gamma_k} - d)(1 - \gamma_k) + 0.(1 - \gamma_i) \geq \phi_i \tag{21}$$

$$\Rightarrow d \leq \frac{\phi_{i_0} - \phi_i}{1 - \gamma_k} \tag{22}$$

$$\Leftrightarrow d \leq \frac{\phi_{i_0} - \max_{i=1, i \neq i_0}^{k-1} \phi_i}{1 - \gamma_k} = d_{max} \tag{23}$$

Similar to the analysis of Proposition 1, considering the expected gain and the risk of the adversary being detected, 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. Therefore, the adversary needs to create at least $\mid D_{ext} \mid = \frac{\phi_{i_0}}{1-\gamma_k}$ identities, and use these identities to post $C_{ext} = \mid D_{ext} \mid = \frac{\phi_{i_0}}{1-\gamma_k}$ ratings on the item s_k.

For the other case where $\gamma_{i_0} < \gamma_k$, finding the solution set is non-trivial.

B Proof for Proposition 3

Proof. Given an item $s'_j \in S2$ from the second system, we define the following similar notations (see Fig. 2(b) for an illustration):

- $U2_j$: all honest users with ratings on s'_j. $U2 = \bigcup_{s'_j \in S2} U2_j$ is thus the set of all honest users with ratings on some items.
- $U2'_j \subseteq U2_j$: honest users with wrong rating on s'_j (due to observation noise)

- $U2''_j \subseteq U2'_j$: honest users with wrong rating on s'_j and detected as cheaters. The set $U2'' = \bigcup_{s_j \in S2} U2''_i$ is the set of honest users of the second system wrongly detected as cheaters.
- $D2_j \subseteq D2$: cheating users with wrong ratings on s'_j (malicious behaviors). $D2 = \bigcup_{s'_j \in S2} D2_j$ is thus the set of malicious users in the second system.
- $L2_j \subseteq D2_j$: cheating users undetected by the trust evaluation in the second system. The set of cheaters detected is $D2_j - L2_j$. Hence $D2 - L2 = \bigcup_{s_j \in S2} (D2_i - L2_i)$ is the set of cheating users correctly identified by the second system.

The set of raters detected as cheating by the second system are therefore in $U2'' \cup (D2 - L2)$. This set is shared by the second system for the first. In exchange, the first system also shares similar result for the second one. Such a sharing can be generalized to many systems, so as the analysis in this work. Fair and reliable sharing of such information between two systems are not considered in this work. We assume that open trust systems are designed to automatically share results of their malicious detection to each other reliably, and system managers have little incentives to modify the software implementation to send wrong information.

The misbehavior detection $U2'' \cup (D2 - L2)$ from the second system helps to identify more cheaters for the first systems. Our main hypothesis is that there are certain users (both honest and malicious) who appear and rate items on both systems. Such common users do exist due to various reasons. For example, the adversary may have items in both systems, whose ranks need to be boosted. As identities are costly, it may need to reuse some identities in one system for another. This is even more true in cases where real persons are hired by (bad) providers to rate and promote their items, more malicious behaviors on many items on different systems bring more benefits to users.

For any user $u \in U \cup D$, any item $s_i \in S, i = 1, ..., k$, the first system uses a modified trust evaluation function $\widehat{t}(u, s_i)$ for every rating $r(u, s_i)$ as follows.

1. $\widehat{t}(u, s_i) = 0$ if $\exists v \in U2'' \cup (D2 - L2)$ and u and v can be linked to the same user, e.g., using the same credentials such as email when registration. Thus any user detected (wrongly or correctly) as cheating in the second system and appears in the first system are also marked as cheating by the first[4]. Concretely, $\widehat{t}(u, s_i) = 0$ for $u \in (D_i - L_i) \cup U''_i \cup (L_i \cap (D2 - L2)) \cup ((U_i - U''_i) \cap U2'')$ (the shaded parts in Fig. 2(b)).
2. For other ratings $r(u, s_i)$ where u does not appear and detected as cheating in the second system, the trustworthiness of the rating is define according to the majority rule. That is, define $P_i = U_i - U'_i - ((U_i - U'_i) \cap U2'')$ as the group of users voting positively on s_i, $N_i = [L_i - (L_i \cap (D2 - L2))] \cup [(U'_i - U''_i) - ((U'_i - U''_i) \cap U''_2)]$ the group of users voting negatively on s_i, we have:
   - For $u \in P_i : \widehat{t}(u, s_i) =\mid P_i \mid /(\mid P_i \mid + \mid N_i \mid)$.
   - For $u \in N_i : \widehat{t}(u, s_i) =\mid N_i \mid /(\mid P_i \mid + \mid N_i \mid)$.

From section 3, we have the following observations:

[4] Different trust integrating policies can be used here, but we limit our study to only this simple case

- $\mid U_i \mid = x_i$, $E \mid U_i' \mid = x_i\varepsilon$, $E \mid U_i'' \mid = x_i\varepsilon\gamma$, so $E \mid U_i - U_i'' \mid = x_i(1-\varepsilon\gamma)$.
- $\mid D_i \mid = y_i$, $E \mid L_i \mid = y_i(1-\gamma)$.

Let $z_i = \mid D_i \cap D2 \mid \leq y_i, 1 \leq i \leq k$ be the number of cheating raters who posts biased ratings on the item s_i of the first system and who also appears in the second system. Denote $\delta_2(j)$ the probability a cheating user u rates on an item $s_j \in S2$, we have:

$$\begin{aligned}
E[\mid L_i \cap (D2 - L2) \mid] &= E[\sum_{u \in D_i \cap D2} 1_{\{u \text{ rates } s_i \in S \text{ and not detected}\}} 1_{\{u \text{ rates } s_j \in S2 \text{ and detected}\}}] \\
&= \mid D_i \cap D2 \mid Pr[u \text{ rates on } s_i \in S \text{ and not detected}] \times \\
&\quad Pr[u \text{ rates on } s_j \in S2 \text{ and detected}] \\
&= z_i(1-\gamma) \sum_{s_j \in S2} \gamma_2 \delta_2(j) \\
&= z_i(1-\gamma)\gamma_2
\end{aligned}$$

Note that $\tau_i = \mid U_i \cap U2 \mid, 1 \leq i \leq k$:

$$\begin{aligned}
E \mid U_i - U_i' \cap U2'' \mid &= E[\sum_{u \in U_i \cap U2} 1_{\{u \text{ rates correctly on } s_i \in S\}} \times \\
&\qquad 1_{\{u \text{ rates wrongly on } s_j \in S2 \text{ and incorrectly detected as cheater}\}}] \\
&= \mid U_i \cap U2 \mid Pr[u \text{ rates correctly on } s_i \in S] \times \\
&\quad Pr[u \text{ rates wrongly on } s_j \in S2 \text{ and incorrectly detected as cheater}] \\
&= \tau_i(1-\varepsilon)\varepsilon\gamma_2
\end{aligned}$$

The expected sizes of the two groups P_i and N_i are:

$$\begin{aligned}
E[\mid P_i \mid] &= E[\mid (U_i - U_i') - ((U_i - U_i') \cap U2'') \mid] \\
&= E[\mid U_i - U_i' \mid] - E[\mid (U_i - U_i') \cap U2'' \mid] \\
&= x_i(1-\varepsilon) - \tau_i(1-\epsilon)\epsilon\gamma_2 \\
&= (x_i - \tau_i\epsilon\gamma_2)(1-\varepsilon) \\
E[\mid N_i \mid] &= E[\mid L_i - (L_i \cap (D2 - L2)) \mid] + E[\mid L_i \mid - E \mid L_i \cap (D2 - L2) \mid] \\
&= y_i(1-\gamma) - z_i(1-\gamma)\gamma_2 + (x_i\varepsilon - x_i\varepsilon\gamma) - \tau_i\varepsilon(1-\gamma)\varepsilon\gamma_2 \\
&= y_i(1-\gamma) - z_i(1-\gamma)\gamma_2 + \varepsilon(1-\gamma)(x_i - \tau_i\varepsilon\gamma_2)
\end{aligned}$$

Similar to the case of one system (Section 3), one may verify that the trust-based QP-score of an item $s_i \in S, i = 1, ..., k-1$ can be computed as:

$$\begin{aligned}
f_T(s_i) &= \mid P_i \mid - \mid N_i \mid \\
\Rightarrow E[f_T(s_i)] &= E \mid P_i \mid - E \mid N_i \mid \\
&= (x_i - \tau_i\epsilon\gamma_2)(1-\varepsilon) - y_i(1-\gamma) + z_i(1-\gamma)\gamma_2 - \varepsilon(1-\gamma)(x_i - \tau_i\varepsilon\gamma_2) \\
&= (x_i - \tau_i\varepsilon\gamma_2)(1 - 2\varepsilon + \varepsilon\gamma) - y_i(1-\gamma) + z_i(1-\gamma)\gamma_2
\end{aligned}$$

Similarly, the modified score of the target item s_k is (noting that honest users mostly vote negatively on s_k):

$$\begin{aligned}
E[f_T(s_k)] &= -E \mid P_k \mid + E \mid N_k \mid \\
&= -(x_k - \tau_k\epsilon\gamma_2)(1 - 2\varepsilon + \varepsilon\gamma) + y_k(1-\gamma) - z_k(1-\gamma)\gamma_2
\end{aligned}$$

In expectation, the item s_k is ranked higher than s_i iff $E[f_T(s_k)] \geq E[f_T(s_i)]$. In other words:

$$y_k + y_i \geq [x_k + x_i - \epsilon\gamma_2(\tau_k + \tau_i)]\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} + (z_k + z_i)\gamma_2$$

The optimal number of ratings $C_{\widehat{T}}$ of the adversary to successfully attack the first systems is the solution to the following linear integer program (IP):

$$C_{\widehat{T}} = \min\{y_1 + y_2 + ... + y_k\} \text{ subject to:}$$

$$y_k + y_i \geq (x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} + (z_k + z_i)\gamma_2, i = 1, ..., k - 1 \quad (24)$$

$$y_i \geq z_i, i = 1, ..., k \quad (25)$$

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 $i \leq j, i, j = 1, ..., k - 1$.

Use the new variables $u_i = y_i - z_i, i = 1, ..., k$, we have another equivalent program (IP_2):

$$\begin{aligned} &\min\{u_1 + u_2 + ... + u_k\} \text{ subject to:} \\ &u_k + u_i \geq g_i, i = 1, ..., k - 1 \\ &u_i \geq 0, i = 1, ..., k \end{aligned} \quad (26)$$

where $g_i, i = 1, ...k$ is defined as:

$$g_i = (x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} + (z_k + z_i)\gamma_2 - z_i - z_k \quad (27)$$

$$= (x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} - (z_k + z_i)(1 - \gamma_2) \quad (28)$$

Suppose that $\widehat{u}_i, i = 1, ..k$ be the solution for the IP_2. From (26), let $i_0 = argmax_{i=1}^{k-1} g_i$, then:

$$\widehat{u}_k + \widehat{u}_{i_0} \geq \max_{i=1}^{k-1} g_i \quad (29)$$

Since $u_i \geq 0, i = 1, ..., k$, the complete set of solutions to IP_2 is:

$$\widehat{u}_k = \max\{0, \max_{i=1}^{k-1} g_i\} - d \quad (30)$$

$$\widehat{u}_{i_0} = d \quad (31)$$

$$\widehat{u}_i = 0, \text{ for } 1 \leq i \leq k - 1, i \neq i_0 \quad (32)$$

$$\text{where} \quad 0 \leq d \leq d_{max} = \max\{0, \max_{i=1}^{k-1} g_i\} - \max\{0, \max_{i=1, i\neq i_0}^{k-1} g_i\} \quad (33)$$

The solution of the original program IP is thus:

$$\widehat{y}_k = \max\{0, \max_{i=1}^{k-1} g_i\} + z_k - d \quad (34)$$

$$\widehat{y}_{i_0} = z_{i_0} + d, \quad (35)$$

$$\widehat{y}_i = z_i, 1 \leq i \leq k - 1, i \neq i_0 \quad (36)$$

$$\text{where} \quad 0 \leq d \leq d_{max} = \max\{0, \max_{i=1}^{k-1} g_i\} - \max\{0, \max_{i=1, i\neq i_0}^{k-1} g_i\} \quad (37)$$

Any of the above solutions requires the adversary to post the same optimal number of ratings:

$$C_{\hat{T}} = \sum_{i=1}^{k} \widehat{y}_i = \max\{0, \max_{i=1}^{k-1} g_i\} + \sum_{i=1}^{k} z_i \quad (38)$$

$$= \max\{0, \max_{i=1}^{k-1}\{(x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} - (1 - \gamma_2)(z_k + z_i)\}\} + \sum_{i=1}^{k} z_i \quad (39)$$

Similar to the previous section, we assume that the adversary cares most about the probability of success of the attack. The optimal strategy of the adversary is for $d = 0$, otherwise with the same user posting ratings on two items s_{i_0} and s_k, the probability that these identities are detected would be higher. This optimal strategy means that the adversary creates and uses all $\mid D_{\widehat{T}} \mid = C_{\widehat{T}}$ identities to post positive ratings on s_k. In fact, we can extend the analysis to the case where the optimal strategy is for $d > 0$, for which the adversary needs to create a smaller number of identities. However, the analysis becomes much more complex.

The reason the adversary uses malicious users with similar identities in the two systems for the sake of saving its cost when attacking both systems. Hence it is necessary to estimate whether the adversary should use common malicious users in two systems to minimize the total number of identities to be created.

Consider the first system. The optimal solution $\widehat{y}_i, i = 1, .., k$ given in (34,35, 36) corresponds to the following strategy of the adversary to boost rank of the item s_k.

- For the target item s_k: the adversary uses a set of malicious 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 $\mid D_k \mid = \max\{0, \max_{i=1}^{k-1} g_i\}$.
- For 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.

Therefore, the set of malicious users to be used by the adversary in the first system is:

$$D = D_k \cup Z \quad (40)$$

where $Z \subseteq D2$ is set of identities borrowed from the set of malicious users $D2$ in the second system. These borrow identities are used by the adversary to post a total of $\sum_{i=1}^{k} z_i$ ratings on those items $s_i, i = 1, ..., k$.

Assume 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 1. By similar reasons, the set of malicious users in the second system is:

$$D2 = D_{k_2} \cup Z2 \quad (41)$$

$Z2 \subseteq D$ 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 only appear in the second system and is used to rate the item $s'_{k_2} \in S2$.

The set of malicious users used by the adversary to attack both systems is:

$$D_{\widehat{T}} = D_k \cup Z \cup D_{k_2} \cup Z2 \tag{42}$$

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, e.g., to rate item $s_i, i = 1, ..., k$ in the first system. Under such a situation, the total minimal number of identities the adversary needs to create in the two systems is:

$$\mid D_{\widehat{T}} \mid = \mid D_k \mid + \mid D_{k_2} \mid \tag{43}$$

From Eq. (28):

$$\mid D_k \mid = \max\{0, \max_{i=1}^{k-1} g_i\} \tag{44}$$

$$= \max\{0, \max_{i=1}^{k-1}\{(x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma} - (z_k + z_i)(1 - \gamma_2)\}\} \tag{45}$$

Similarly, we have:

$$\mid D_{k_2} \mid = \max\{0, \max_{1 \le i \le k_2 - 1}\{(x'_{k_2} + x'_i - \varepsilon\gamma(\tau'_k + \tau'_i))\frac{1 - \varepsilon + 2\varepsilon\gamma_2}{1 - \gamma_2} - (z'_k + z'_i)(1 - \gamma)\}\} \tag{46}$$

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

The optimal number of ratings $C_{\widehat{T}}$ the adversary must post in the first system, given fixed $z_i, i = 1, ..., k$ is:

$$C_{\widehat{T}} = \max\{0, \max_{i=1}^{k-1}\{(x_k + x_i - 2\varepsilon + \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 1\varepsilon + \varepsilon\gamma}{1 - \gamma} - (1 - \gamma_2)(z_k + z_i)\}\} + \sum_{i=1}^{k} z_i \tag{47}$$

$$= \mid D_k \mid + \sum_{i=1}^{k} z_i \tag{48}$$

Likewise, the optimal number of ratings $C'_{\widehat{T}}$ 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 $z_i \ge 0, z'_j \ge 0, i = 1, ..., k, j = 1, ..., k_2$ such that:

$$\text{minimize } R_{\widehat{T}} = \mid D_k \mid + \mid D_{k_2} \mid + \sum_{i=1}^{k} z_i + \sum_{j=1}^{k_2} z'_j \tag{49}$$

$$= \mid D_{\widehat{T}} \mid + \sum_{i=1}^{k} z_i + \sum_{j=1}^{k_2} z'_j \tag{50}$$

$$\text{subject to: } \mid D_{\widehat{T}} \mid = \mid D_k \mid + \mid D_{k_2} \mid = N \tag{51}$$

Or equivalently:

$$\text{minimize } R_{\widehat{T}} = N + \sum_{i=1}^{k} z_i + \sum_{j=1}^{k_2} z'_j \tag{52}$$

$$\text{subject to: } \mid D_k \mid + \mid D_{k_2} \mid = N \tag{53}$$

For simplicity, denote:

$$f_i \triangleq (x_k + x_i - \varepsilon\gamma_2(\tau_k + \tau_i))\frac{1 - 2\varepsilon + \varepsilon\gamma}{1 - \gamma}, i = 1, ..., k-1 \tag{54}$$

$$f'_i \triangleq (x'_{k_2} + x'_i - \varepsilon\gamma(\tau'_{k_2} + \tau'_i))\frac{1 - 2\varepsilon + \varepsilon\gamma_2}{1 - \gamma_2}, i = 1, ..., k_2 - 1 \tag{55}$$

We can find the solution of the above optimization problem as follow. Assume $\mid D_k \mid = n < N$, and $\mid D_{k_2} \mid = N - n$ in Eq. (53). From Eq. (45,46), we have:

$$\mid D_k \mid = \max\{0, \max_{i=1}^{k-1}\{f_i - (z_k + z_i)(1 - \gamma_2)\}\} = n \tag{56}$$

$$\Rightarrow f_i - (z_k + z_i)(1 - \gamma_2) \leq n, \forall i = 1, ..., k-1$$

$$\Rightarrow z_k + z_i \geq \frac{f_i - n}{1 - \gamma_2}, \forall i = 1, ..., k-1 \tag{57}$$

$$\Rightarrow \exists i_0 \in [1, k-1] : z_k + z_{i_0} \geq \max_{i=1}^{k-1} \frac{f_i - n}{1 - \gamma_2} \tag{58}$$

For $n \geq 0$ and $\widehat{z}_k \geq 0$, we note that $n \leq \max_{i=1}^{k-1} f_i < N$.

From the constraint (58), it follows that $\sum_{i=1}^{k} z_i \geq \max_{i=1}^{k-1} \frac{f_i - n}{1 - \gamma_2}$. The minimal value $\sum_{i=1}^{k} z_i = \max_{i=1}^{k-1} \frac{f_i - n}{1 - \gamma_2}$ is attained for any $0 \leq n \leq N$, at for example, $\widehat{z}_k = \max_{i=1}^{k-1} \frac{f_i - n}{1 - \gamma_2}, \widehat{z}_i = 0, i = 1, ..., k-1$.

Due to the symmetry, we also get $\sum_{j=1}^{k_2} z'_j \geq \max_{1 \leq j \leq k_2 - 1} \frac{f'_j - (N-n)}{1 - \gamma}$. One possible assignment is $z'_{k_2} = \widehat{z'}_{k_2} = \max_{1 \leq j \leq k_2 - 1} \frac{f'_j - (N-n)}{1 - \gamma}$, and $z'_j = \widehat{z'}_j = 0, j = 1, ..., k_2 - 1$ for any $0 \leq n \leq N$.

The objective function in Eq. (52) gives us:

$$\begin{aligned} R_{\widehat{T}} &\geq N + \sum_{i=1}^{k} \widehat{z}_i + \sum_{j=1}^{k_2} \widehat{z}'_j \\ &\geq N + \max_{i=1}^{k-1}\{\frac{f_i}{1 - \gamma_2}\} - \frac{n}{1 - \gamma_2} + \max_{j=1}^{k_2 - 1}\{\frac{f'_j}{1 - \gamma}\} - \frac{N - n}{1 - \gamma} \\ &\geq N + \max_{i=1}^{k-1}\{\frac{f_i}{1 - \gamma_2}\} + \max_{j=1}^{k_2 - 1}\{\frac{f'_j}{1 - \gamma}\} - (\frac{n}{1 - \gamma_2} + \frac{N - n}{1 - \gamma}), \forall 0 \leq n \leq N \end{aligned} \tag{59}$$

We consider the two following cases:

- $\gamma = \gamma_2$, i.e., the two systems have comparable probability of detecting misbehavior. In this case, one can verify that:

$$R_{\widehat{T}} \geq -\frac{N\gamma}{1 - \gamma} + \max_{i=1}^{k-1}\{\frac{f_i}{1 - \gamma}\} + \max_{j=1}^{k_2 - 1}\{\frac{f'_j}{1 - \gamma}\} \tag{60}$$

- $\gamma \neq \gamma_2$: without loss of generality, assume that $\gamma_2 > \gamma$, or the second system offer better capability of detecting malicious ratings than the first one. Thus $\frac{n}{1-\gamma_2} + \frac{N-n}{1-\gamma} = \frac{N}{1-\gamma} + n(\frac{1}{1-\gamma_2} - \frac{1}{1-\gamma})$ is maximized at the maximal value of $n = \max_{i=1}^{k-1} f_i$. In this case, we have $\widehat{z}_i = 0, i = 1, ..., k$, or the adversary uses no identities from the second system. From (59), the optimal number of ratings in this case is:

$$R_{\widehat{T}} \geq \max_{i=1}^{k-1}\{\frac{f_i}{1-\gamma_2}\} + \max_{j=1}^{k_2-1}\{\frac{f'_j}{1-\gamma}\} \tag{61}$$

$$+N - (\frac{N}{1-\gamma} + \max_{i=1}^{k-1} f_i(\frac{1}{1-\gamma_2} - \frac{1}{1-\gamma})) \tag{62}$$

$$\geq -\frac{N\gamma}{1-\gamma} + \max_{i=1}^{k-1}\{\frac{f_i}{1-\gamma}\} + \max_{j=1}^{k_2-1}\{\frac{f'_j}{1-\gamma}\} \tag{63}$$

A difference between the two cost is computed as follows, given a known Δ.

$$\begin{aligned} R_{\widehat{T}} - C_T - C'_T &> -\frac{N\gamma}{1-\gamma} + \max_{i=1}^{k-1}\{\frac{f_i}{1-\gamma}\} + \max_{j=1}^{k_2-1}\{\frac{f'_j}{1-\gamma}\} - C_T - C'_T \\ &= -\frac{(C_T + C'_T - \Delta)\gamma}{1-\gamma} - C_T - C'_T + \max_{i=1}^{k-1}\{\frac{f_i}{1-\gamma}\} + \max_{j=1}^{k_2-1}\{\frac{f'_j}{1-\gamma}\} \\ &> \frac{\Delta\gamma}{1-\gamma} + \max_{i=1}^{k-1}\{\frac{f_i}{1-\gamma}\} - \frac{C_T}{1-\gamma} + \max_{j=1}^{k_2-1}\{\frac{f'_j}{1-\gamma}\} - \frac{C'_T}{1-\gamma} \end{aligned} \tag{64}$$

On the other hand, from Eq (54), and note that $0 < \gamma_2 < 1$, $\tau_i \leq x_i, i = 1, ..., k$, one has:

$$\begin{aligned} \max_{i=1}^{k-1}\{\frac{f_i}{1-\gamma}\} - \frac{C_T}{1-\gamma} &\geq \frac{f_1}{1-\gamma} - \frac{C_T}{1-\gamma} \\ &\geq \frac{1}{1-\gamma}([x_k + x_1 - \epsilon\gamma_2(\tau_k + \tau_1)]\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma} - (x_k + x_1)\frac{1-2\varepsilon++\varepsilon\gamma}{1-\gamma}) \\ &= -\frac{\varepsilon\gamma_2(\tau_k + \tau_1)(1-2\varepsilon+\varepsilon\gamma)}{(1-\gamma)^2} \end{aligned} \tag{65}$$

By symmetry, we also have:

$$\max_{j=1}^{k_2-1}\{\frac{f'_j}{1-\gamma}\} - \frac{C'_T}{1-\gamma} \geq -\frac{\varepsilon\gamma(\tau'_{k_2} + \tau'_1)(1-2\varepsilon+\varepsilon\gamma_2)}{(1-\gamma_2)^2} \tag{66}$$

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:

$$0 \leq \Delta \leq \min\{C_T, C'_T\} = \min\{(x_k + x_1)\frac{1-2\varepsilon+\varepsilon\gamma}{1-\gamma}, (x'_{k_2} + x'_1)\frac{1-2\varepsilon+\varepsilon\gamma_2}{1-\gamma_2}\} \tag{67}$$

With Δ defined as above, and from the two inequalities (65,66) in (64), it follows that:

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

and Proposition 3 follows naturally. □

References

[1] Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research **43** (2006)

[2] Parsa, A.: Belkins Development Rep is Hiring People to Write Fake Positive Amazon Reviews. (2009)

[3] 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** (2009) 1422–1438

[4] 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). (2004) 271–279

[5] Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University (1998)

[6] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46** (1999) 604–632

[7] Golbeck, J.: Trust on the world wide web: A survey. Foundations and Trends in Web Science **1** (2006) 131–197

[8] Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decis. Support Syst. **43** (2007) 618–644

[9] 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 (2010)

[10] Vu, L.H., Hauswirth, M., Aberer, K.: Qos-based service selection and ranking with trust and reputation management. In: International Conference on Cooperative Information Systems (CoopIS). (2005)

[11] Kamvar, S.D., Schlosser, M.T., Molina, H.G.: The EigenTrust algorithm for reputation management in P2P networks. In: Proc. of WWW'03. (2003)