Online Appendix to: Effective Usage of Computational Trust Models in Rational Environments

LE-HUNG VU and KARL ABERER, Ecole Polytechnique Fédérale de Lausanne (EPFL)

A. EXAMPLES OF THE BASIC CONCEPTS

Example A.1. The personalized trust model PeerTrust [Xiong and Liu 2004] can be used to evaluate rating reliability via the following formalism.

— $P_i = S(i) \cap S(j)$. In other words, the set of relevant peers P_i includes those peers k having provided services to both i, j.

— $V_i = \{r(x, k, tr) \mid x \in \{i, j\}, k \in P_i, tr \in \Omega(i, k) \cup \Omega(j, k)\}$, that is, V_i consists of those ratings by i and j on service behaviors of other peers k in the relevant set P_i.

— The relationship W among peers and the features of the target $\mathcal{F}_j$ are not considered.

— $T_{ij} = 1 - \sqrt{\frac{\sum_{k \in P_i} \frac{\sum_{t_{ik} \in \Omega(i,k)} r(i,k,t_{ik})}{\|\Omega(i,k)\|} - \frac{\sum_{t_{jk} \in \Omega(j,k)} r(j,k,t_{jk})}{\|\Omega(j,k)\|}}{\|P_i\|}}$ is a measure of similarity between the possible ratings $r(i, k, t_{ik})$ and $r(j, k, t_{ik})$ on those transactions $t_{ik} \in \Omega(i, k)$ and $t_{jk} \in \Omega(j, k)$.

— A rating by j is considered as reliable if $T_{ij} > T_{min}$ and as unreliable otherwise, where T_{min} is a possibly global system design threshold. Alternatively, we can normalize T_{ij} into [0, 1] and trust the rating with probability T_{ij}. With probability $1-T_{ij}$, the rating is distrusted (evaluated as unreliable).

Example A.2. Another approach to estimate a peer's trustworthiness is to assume that peers behave according to a probabilistic model. Similarity in rating on one target leads to similarity in rating on another [Vu and Aberer 2007]. The peer i estimates that the target peer j has a probability T_{ij} of reporting truthfully what j observes. This model is specified similar to Example A.1, except that T_{ij} D_i are defined differently.

$$T_{ij} = \frac{\sum_{k \in P_i, t_{ik} \in \Omega(i,k), t_{jk} \in \Omega(j,k)} I(r(i, k, t_{ik}) = r(j, k, t_{jk}))}{\|P_i\| \sum_{k \in P_i} \|\Omega(i, k)\| \sum_{k \in P_i} \|\Omega(j, k)\|}, \tag{2}$$

where the indicator function $I(c)$ evaluates to 1 if the Boolean condition c is true. Thus T_{ij} is defined by the fraction of ratings by j having the same values as ratings by i. This T_{ij} is an estimate of the probability of the peer j being honest when rating, and such an estimate maximizes the likelihood of having the observation set V_i by the set of relevant peers P_i. The decision rule D_i is usually probabilistic, e.g., trust the rating with probability T_{ij} and distrust it with probability $1 - T_{ij}$.

Example A.3. It can be shown that the naive computational trust model $\mathcal{N}$, which trusts any rating and considers no rating as the presence of a positive one, has the misclassification errors $\alpha = \alpha_0 = 1$ and $\beta = \beta_0 = 0$. In fact, let $0 \leq h, l, i \leq 1$, where $h + l + i = 1$ be respectively the probabilities that the rating peer provides a reliable rating, an unreliable one, and no rating after a transaction with a specific provider. Denote as *est*+ (resp. *est*−) the events that the most recent rating is evaluated by the learning peer as reliable (resp. unreliable), and let *real*+ (resp. *real*−) be the events that the rating is actually reliable (resp. unreliable). There are two possibilities.

DOI 10.1145/2019591.2019593 http://doi.acm.org/10.1145/2019591.2019593

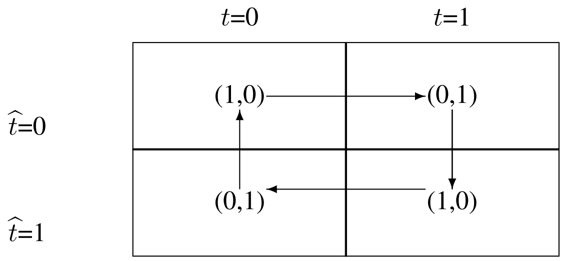

Fig. 11. The game between a learning peer (the row player) and the strategic rater (the column player). The notation (x, y) means that the payoff of the learning peer is x and of the rater is y. Since the goal of a rational rater is to maximize misclassification errors of the dishonesty detector, we give a rater a payoff 1 if $t \neq \widehat{t}$ and 0 if $t = \widehat{t}$. The payoff of a learning peer is the opposite.

- The provider did cooperate in the last transaction, thus the absence of a rating is equivalent to the presence of a reliable positive rating. So $\alpha_0 = \frac{Pr(est+,real-)}{Pr(real-)} = l/l = 1$ and $\beta_0 = \frac{Pr(est-,real+)}{Pr(real+)} = 0/(h+i) = 0$.
- The provider did not cooperate in the last transaction, thus the absence of a rating implies the presence of an unreliable positive rating. Still, we have $\alpha_0 = \frac{Pr(est+,real-)}{Pr(real-)} = (l+i)/(l+i) = 1$ and $\beta_0 = \frac{Pr(est-,real+)}{Pr(real+)} = 0/h = 0$.

B. PROOF OF PROPOSITION 3.1

PROOF. Let t and $\widehat{t}$ be the binary reliability of a rating, as exhibited by the rater and as estimated by the learning peer, respectively. Denote $\mathcal{H} = \langle P_i, W, V_i, \mathcal{F}_j \rangle$ the input of the algorithm A (c.f. Definition 2.1), we have:

$$\alpha = Pr(\widehat{t} = 1 \mid t = 0, \mathcal{H}) \propto Pr(\widehat{t} = 1, t = 0 \mid \mathcal{H}). \tag{3}$$

If the model $\mathcal{R}$ is publicly known and the decision rules D are deterministic, a rational rater knows exactly whether the rating is estimated as reliable given the history $\mathcal{H}$. Therefore, the rater can strategically provide rating with an opposite reliability to maximize misclassification errors of the dishonesty detector. The game between a learning peer and the rater is shown in Figure 11.

The arrows in Figure 11 denote the possible moves of each player to maximize its payoff, showing that the game has no pure Nash equilibrium. The only mixed equilibrium of the game is: the rater exhibits a random rating strategy $Pr(t = 1) = Pr(t = 0) = 0.5$ and the learning peer estimates the rating reliability as $Pr(\widehat{t} = 1) = Pr(\widehat{t} = 0) = 0.5$, so $\alpha = \beta = 0.5$. □

C. PROOF OF THEOREM 3.2

PROOF. We first prove (1). Rational providers apparently do not find incentives to cooperate in the last transaction. Consider those rational providers staying in the system for $\Delta > 1$ more transactions after the current one.

Let $0 \leq h, s, l, i \leq 1$ respectively, be the probabilities that the current client exhibits following rating behaviors after the transaction: honest (provides reliable ratings), advertising (posts positive ratings on the provider), badmouthing (rates the provider negatively), and nonparticipating (does not leave any rating), where $h + s + l + i = 1$. Note that possible strategic rating manipulations by any raters colluding with the current provider are all considered by these probabilities. For example, consider the case where the provider may use a fake identity to stuff a positive rating with a newer timestamp to hide its cheating in a transaction. In this case, the provider still has additional gain υ in the transaction, and the dishonesty detection is applied on the fake rating whose rater is a client with completely advertising behavior, i.e., $h = i = l = 0, s = 1$.

The probabilities that an honest provider obtains a positive (resp. negative) rating after a transaction are $h^{+} = h + s + i = 1 - l$ (resp. $1 - h^{+}$). The honest provider is blacklisted if either the true positive rating is not accepted by the computational trust model as reliable (with probability β), or the wrong negative rating is accepted as reliable (with probability α). Thus the probability that the provider will be blacklisted by a forthcoming client is: $x_b = h^{+}\beta + (1 - h^{+})\alpha = (1 - l)\beta + l\alpha \leq \varepsilon$, since $0 \leq l \leq 1$ and $0 \leq \alpha \leq \varepsilon, 0 \leq \beta \leq \varepsilon$.

The probability that the provider is globally blacklisted after the current transaction is then $x_b^k \leq \varepsilon^k$. This inequality holds even in the presence of malicious or strategic manipulation of ratings by any raters with different h, l, s, i, provided that misclassification errors α, β of $\mathcal{R}$ are less than ε.

Similarly, if the provider is cheating in this transaction, the probability that it obtains a positive rating is $l^{+} = s + i = 1 - h - l$. With probability $1 - l^{+}$ such a provider receives a negative rating. In this case, the provider will be blacklisted by a future client with probability $y_b = l^{+}(1 - \alpha) + (1 - l^{+})(1 - \beta) = (1 - h - l)(1 - \alpha) + (h + l)(1 - \beta) \geq 1 - \varepsilon$. Thus the probability that the provider is globally blacklisted is $y_b^k \geq (1 - \varepsilon)^k$.

Let U be the current accumulative utilities of a rational provider and u_h be its best (maximized) expected utilities for the remaining time in the system if it is not globally blacklisted after the current transaction. Denote U_{honest} (and U_{cheat}) as the best (maximal) expected life-time utilities of the provider if it is honest (respectively cheating) in the current transaction, it follows that:

$$\begin{aligned} U_{honest} &= U + u + u_h(1 - x_b^k) \\ U_{cheat} &= U + (u + \upsilon) + u_h(1 - y_b^k) \\ \delta_{hc} &= U_{honest} - U_{cheat} = -\upsilon + u_h(y_b^k - x_b^k) \\ &\geq -\upsilon + u_h((1 - \varepsilon)^k - \varepsilon^k). \end{aligned}$$

One can verify that the preceding reasoning is applicable in the following two situations. First, identities are very difficult to obtain and thus the provider cannot rejoin under a new identity. Second, the cost of obtaining a new identity outweighs the maximal temporary benefit gained by cheating in a transaction.

As an honest provider is still blacklisted with probability $x_b^k < \varepsilon^k$, one can verify that the fully cooperative strategy of a provider during $\Delta \geq 1$ transactions leads to a total utility of at least $\frac{1-(1-x_b^k)^\Delta}{x_b^k} u_* \geq \frac{1-(1-\varepsilon^k)^\Delta}{\varepsilon^k} u_* > 0$. Note that for small x_b^k, $u_h \geq \Delta u_*$ approximately.

It follows that $\delta_{hc} \geq 0$ if $\Delta \geq \frac{\ln[1 - \frac{\upsilon\varepsilon^k}{u_*((1-\varepsilon)^k - \varepsilon^k)}]}{\ln(1-\varepsilon^k)} = \Delta_\upsilon$, where $\varepsilon < \varepsilon_{max}(k) = 1/(1 + \sqrt[k]{1 + \upsilon^*/u_*}) < 0.5$ so that the logarithm is always well-defined for any $\upsilon \leq \upsilon^*$.

Therefore, in any transaction but its last Δ_υ ones, a rational provider considers cooperation as its best response strategy. Thus (1) is proven. The proof of (2) is then straightforward from the preceding analysis.

To prove (3), note that after each transaction, the probability that by accident, an honest provider is globally blacklisted is $x_b^k \leq \varepsilon^k$. In the worst case ever, N_h is a geometric random variable with probability ε^k, hence $E[N_h] > 1/\varepsilon^k$.

By similar reasoning, the probability a malicious provider is globally blacklisted is $y_b^k \geq (1 - \varepsilon)^k$, and thus $E[N_c] < 1/(1 - \varepsilon)^k$ □

D. PROOF OF COROLLARY 3.4

PROOF. The naive computational model $\mathcal{N}$ has misclassification errors $\alpha = 1$ and $\beta = 0$ (Example A.3). Proceeding as in the analysis of Theorem 3.2, we have $\delta_{hc} \geq -v^* + u_h(h-(1-h)) = -v^* + u_h(2h-1) \geq -v^* + (1-h^{\Delta})(2h-1)/(1-h)$ (herein $u_h \geq (1-h^{\Delta})/(1-h)$).

A rational provider would cooperate if $\delta_{hc} \geq 0$, or $\Delta \geq \frac{\ln[1-(1-h)/(2h-1)]}{\ln h}$. The condition $1 > h > h_{min} = (1 + v^*/u_*)/(1 + 2v^*/u_*)$ makes the logarithm well-defined.

Note that $\delta_{hc} \geq -v^* + (1 - h^{\Delta})(2h - 1)/(1 - h)$, where the right-hand side is monotonically increasing in h. This fact gives direct incentives for a long-staying client to leave a correct rating after a transaction so as to increase the overall probability of reporting truthfully h of any client as estimated by subsequent providers. This maximizes the chance of this current client to have successful transactions in the future even with other rational providers (for larger h, δ_{hc} gets larger and thus it is more favored for the future provider to cooperate than to cheat). □

E. PROOF OF THEOREM 4.1

PROOF. Let $\delta \geq \Delta$ be the number of remaining transactions of the provider at the current step. Proceed as in Theorem 3.2 with $k = 1$, $\alpha' = c\alpha + (1 - c)\alpha_2 = 1 - c + c\alpha$, and $\beta' = c\beta + (1 - c)\beta_2 = c\beta$, we get $\delta_{hc} \geq -v^* + u_h(h(1 - c) + c(1 - \alpha - \beta - (\beta - \alpha)h))$. Here, the probability that an honest provider is blacklisted is $x_b = (1 - l)\beta' + l\alpha' = c(1 - l)\beta + l(1 - c + c\alpha)$, where l is a small probability that someone badmouths the provider. Thus we have $x_b = (1 - c)l + c[(1 - l)\beta + l\alpha] \leq \max(l, \varepsilon)$, which is small. Therefore, approximately, $u_h \geq \delta u_*$. As a result, $\delta_{hc} \geq -v^* + \delta u_*(h(1 - c) + c(1 - \alpha - \beta - (\beta - \alpha)h))$.

Since $0 \leq h \leq 1$, it follows that $\alpha + \beta + (\beta - \alpha)h \leq \alpha + \beta + \max\{\beta - \alpha, 0\} \leq 2\max\{\beta, \alpha\} \leq 2\varepsilon$. Thus, $h(1 - c) + c(1 - \alpha - \beta - (\beta - \alpha)h) \geq c(1 - 2\varepsilon)$ for $c \in [0, 1]$. This makes $\delta_{hc} \geq -v^* + \delta u_* c(1 - 2\varepsilon)$. Equivalently, $\delta_{hc} \geq 0$, or cooperation is a dominant strategy for the provider if and only if $c \geq c_* = \frac{v^*}{\delta u_*(1-2\varepsilon)}$.

According to Theorem 3.2 (k=1) with small ε, using only algorithm $\mathcal{R}_1$ can ensure cooperation of a rational provider in all transactions but its last Δ ones. That is $\delta_{hc} \geq -v^* + \Delta u_*(1 - 2\varepsilon) \geq 0$, or equivalently $\Delta \geq \frac{v^*}{u_*(1-2\varepsilon)}$. Since $\delta \geq \Delta$, one can verify that $c_* \leq 1$ and thus c_* is a valid probability. □