Strategic Attacks on Trust Models via Bandit Optimization

Taha D. Güneş
t.d.gunes@soton.ac.uk

Long Tran-Thanh
l.tran-thanh@soton.ac.uk

Timothy J. Norman
t.j.norman@soton.ac.uk

Agents, Interaction and Complexity Research Group
Electronics and Computer Science
University of Southampton, UK

Abstract

Trust and reputation systems are designed to mitigate risks associated with decisions to rely upon systems over which there is no direct control. The effectiveness of trust models are typically evaluated against relatively shallow metrics regarding the sophistication of potential attacks. In reality, such systems may be open to strategic attacks, which need to be investigated in-depth if trust model resilience is to be more fully understood. Here, we devise an orchestrated attack strategy for a specific state-of-the-art statistical trust model (HABIT). We evaluate how these intelligent attack strategies can influence predictions of target trustworthiness by this model. Our conjecture is that this approach represents a stronger benchmark for the assessment of trust models in general.

1 Introduction

Interactions among autonomous agents inherently involve the risk of encountering behaviour that is unexpected and, potentially, damaging to the goals of the principal. To this end, models of trust are devised to support agents in making decisions regarding with whom to interact. This decision process may, however, be vulnerable to orchestrated attack from adversaries. Existing trust models are, to a substantial extent, not rigorously evaluated with respect to adversarial attack. The typical argument used with respect to attacks on a trust model is: our model is robust to, say, whitewashing attacks; during evaluation, we introduce a number of agents that perform poorly and then return with a new identity; we then demonstrate that our model is, by some metric, robust to this attack [BNS10]. The assumption is that we have a relatively small number (with respect to the population) of agents that have a pretty dumb attack strategy. What is more of a concern for computational trust models is an intelligent, strategic attacker with a specific aim in mind, potentially generating a high pay-off. To our knowledge, there is very little research that has explored this kind of attack, and the potential impact that it may have on computational models of trust.

All the leading statistical models of trust [TLRJ12, RPC06, FBZ13, BNS10, KSGM03] have been evaluated using either (adapted) real-world datasets or via simulation. The metrics of evaluation typically include the *mean absolute error* of an assessment of trustworthiness, or *precision* and *recall* of a classification mechanism

In: R. Cohen, M. Sensoy, and T. J. Norman (eds.): Proceedings of the 20th International Workshop on Trust in Agent Societies, Stockholm, Sweden on July 14-15, 2018, published at http://ceur-ws.org

with assessment against a set of straightforward attacks. As argued by Jøsang [Jøs12, JG09], however, this type of analysis requires additional investigation to ensure robustness in real-world scenarios:

> "Many studies on robustness in the research literature suffer from the authors desire to put their own trust and reputation system (TRS) designs in a positive light, with the result that the robustness analyses often are too superficial and fail to consider realistic attacks. Publications providing comprehensive robustness analyses are rare."

The small number of studies that attempt to investigate the robustness of trust models include Kerr & Cohen [KC09], Wang *et al.* [WMLZ14], Muller *et. al.* [MWLZ16], Bidgoly & Ladani [BL16] and Ruan & Durresi [RD16]. The attacks considered in these studies are rule-based, which are not necessarily representative of what a sophisticated attacker is capable of. Kerr & Cohen, for example, employs a *playbook* approach, where the attacker selects strategies using a similarity metric to assess situations in which engineered strategies have been defined as relevant [KC09]. Wang *et. al.* define a stochastic process to model the attacker's strategy in selecting from representative attack actions [WMZL15].

We take a different approach, using the trust model employed by the decision maker (the principal) to assess service providers given available evidence. A malicious service provider (the attacker), targeting a decision maker, intends to influence their reasoning so that assessments of service providers are more to attacker's benefit. From the attacker's perspective, therefore, the goal is to find what kinds of evidence have most influence on their assessments. This may involve *self-promoting* the attacker, or *slandering* competitors [HZNR09], but we make no prior assumptions about what strategies are most useful. In essence, the aim is to find sets of actions that can affect the evidence used by the principal to shift the trust system to promote the malicious motive.

We ground our attack model on a hierarchical optimistic bandit optimization mechanism, presenting this algorithm in Section 3. This is used to strategically sample the space of possible attacks on a trust model, where we cannot assume that the attack space is convex. In Section 4, we assess the performance of this model against HABIT [TLRJ12], a generic model that has been shown to be highly effective in challenging scenarios and benchmarked against other leading algorithms, and discuss our findings in Section 5. Before giving an overview of our approach, however, we formulate the trust problem itself and show how HABIT exploits evidence available to make trust decisions.

2 Trust Model

Trust assessment is the problem of producing a numeric rating (or ranking) for each of a set of agents, $\mathcal{A} = \{1, \ldots, n\}$, which are assumed to be functionally equivalent. In other words, we do not take into account questions of risk in making delegation decisions, we focus on prediction of performance on some problem. Different models use different kinds of evidence in making these assessment, the most simple being a determination of whether the provision of a service was successful, or a failure. Other models use quality ratings, or take into account multiple dimensions of performance such as quality and time to delivery. In general, however, these are observations made by the principal, tr, of the agent concerned (often referred to as the trustee), te over some period of time (e.g. from time 0 to time t): $O^{0:t}_{tr \to te}$. We assume at each time step an interaction is occurred. All the information that is, in principle, available to form assessments is: $\mathcal{E} = \{O^{0:t}_{i \to j} \mid (i,j) \in \mathcal{R}\}$, where $\mathcal{R}$ is the all agent pairs (i,j). We, therefore, assume, in common with most trust models, that these observations are made by a single principal of the performance of a single agent. Generally goal of a statistical trust model is to estimate the expectation of $p(O_{tr \to te} \mid \mathcal{E})$, which is the observation that the principal, tr, *expects* in a future interaction with te, given the evidence. Of course, the principal may have limited memory of past encounters, or it may have limited, biased or no access to others' observations (reports of others' direct observations).

The HABIT model, proposed by Teacy *et al.* [TLRJ12], is a generic statistical trust model, which is compatible with both continuous and discrete forms of information and claims (evidence) that has been empirically demonstrated to be robust to malicious and noisy information. The model can be instantiated with different types of distribution so that it can consume different types of observation; e.g. success/failure, or a qualitative performance rating. The generic model assumes that the reported performance of each trustee (reputation), $\theta_{.\to te}$, is dependent on a hyper-parameter vector ϕ, which is intended to account for the group behaviour of all trustees. Thus, $\theta_{.\to te}$ is a set that consists of parameter vectors each every principal with a specific trustee, $\theta_{.\to te} = \{\theta_{i \to te} \mid i \in \mathcal{A}\}$. These parameter vectors are used to define the probability of interacting with trustee j with exactly the same experience as that reported from the perspective of principal i will be successfully,

$P(O_{i\to j} = success) = \theta_{i\to j}^{success}$. The aim is to calculate the trustworthiness of trustee te, from the perspective of truster tr by all evidence $\mathcal{E}$, which is $\mathbf{E}[\theta_{tr\to te}|\mathcal{E}]$.

Here, we adopt Dirichlet Process [Fer73] instance of the HABIT model. In this instance, trust assessments are calculated analytically rather than through variational inference or Monte Carlo methods. Given all evidence $\mathcal{E}$, the trustworthiness of trustee te is calculated as:

$$\begin{gathered}\theta_{tr\to te} \sim \mathbf{Dir}(\alpha_{tr\to te}) \\ O_{tr\to te} \sim \mathbf{Cat}(\theta_{tr\to te})^1 \\ \mathbb{E}[\theta_{tr\to te}|\mathcal{E}] = c_0 \frac{\text{Beta}(O^{0:t}_{tr\to te})}{\text{Beta}(\emptyset)} \mathbb{E}[\text{Dir}(\theta_{tr\to te}|O^{0:t}_{tr\to te})] + \sum_{j=1}^{n} \frac{\text{Beta}(O^{0:t}_{tr\to j} \cup O^{0:t}_{tr\to te})}{\text{Beta}(O^{0:t}_{tr\to te})} \mathbb{E}[\text{Dir}(\theta_{tr\to te}|O^{0:t}_{tr\to j} \cup O^{0:t}_{tr\to te})]\end{gathered} \tag{1}$$

where:

$$\text{Dir}(\theta_{i\to j}|O_{i\to j}) = \frac{\prod_{l=1}^{k} (\theta^{(l)}_{i\to j})^{\alpha_l + n_l}}{\text{Beta}(\alpha + n)} \tag{2}$$

is the posterior Dirichlet distribution given the evidence, which increments each α_l by number of times, n_l.

$$\text{Beta}(\alpha) = \frac{\prod_{l=1}^{k} \Gamma(\alpha_l)}{\Gamma(\sum_{l=1}^{k} \alpha_l)} \tag{3}$$

is the multivariate Beta function and c_0 is the chosen Dirichlet process constant, which gives weight to direct observation information $O^{0:t}_{tr\to te}$ more than other observations. Lastly, expectation of a Dirichlet distribution is defined by hyper-parameters, $\alpha = \langle \alpha_1, ..., \alpha_k \rangle$ as:

$$\mathbf{E}[\text{Dir}(\theta_{tr\to te} = l|O_{i\to j})] = \frac{\alpha_l + n_l}{\sum_{i=1}^{k} \alpha_i + n_i} \tag{4}$$

For this investigation, since the evaluation of each interaction is binary, number of values is $k = 2$. The idea behind Equation 2 is the assumption that there is a correlation between the truster's interest, trustee te with other trustees. The equation is taking account of this similarity between trustees.

3 Attacker Model

To misguide a truster, an attacker introduces new information to the system or alter previous historical information. We denote an alteration of the evidence as $\mathcal{E}'$. However for an attacker, the objective is to find the most rewarding modification $\mathcal{E}^* \in \mathcal{X}$ in an attack space $\mathcal{X} = \{\mathcal{E}'_0, ..., \mathcal{E}'_{|\mathcal{X}|}\}$, which is:

$$\mathcal{E}^* = \arg\max_{\mathcal{E}'} \underbrace{\mathbb{E}[\theta_{tr\to te}|\mathcal{E}'] - \mathbb{E}[\theta_{tr\to te}|\mathcal{E}]}_{\text{trust gain}} \tag{5}$$

An attack is an alteration in the evidence available to a principal for misguiding the predictions of the principal about the trustworthiness of other parties. Formally, we define an attack as: $\mathcal{E}' = \{A(O^t_{i\to j})|\ (i,j) \in \mathcal{R}$ and $t' \in \{0, ..., t\}\}$, where $A : \{success, failure\} \to \{success, failure\}$ is the manipulation function. Finding the most rewarding modification $\mathcal{E}^*$ in brute force manner is not feasible, since the search space increases significantly by the number of trustees and the interaction representation (e.g. having a larger k.) If we assume full control over the evidence, the number of possible attacks are:

$$|\mathcal{X}| = \binom{\rho + (|O^{0:t}_{tr\to .}| \cdot k) - 1}{(|O^{0:t}_{tr\to .}| \cdot k) - 1} \tag{6}$$

by weak compositions of ρ [HM04] where ρ is attack power: i.e. the number of observations that will be added to the evidence by the attack. For instance, when the trust system has 20 service providers $|O^{0:t}_{tr\to .}| = 20$ and binary opinions $k = 2$, the number of attacks with respect to the attack power:

[1] $\mathbf{Cat}(\theta_{tr\to te})$ is a categorical distribution, such that $p(O_{tr\to te}) = \theta_{tr\to te}$.

Power ρ	Space size $\lvert\mathcal{X}\rvert$	Power ρ	Space size $\lvert\mathcal{X}\rvert$	Power ρ	Space size $\lvert\mathcal{X}\rvert$
1	38	4	101270	7	38320568
2	741	5	850668	8	215553195
3	9880	6	6096454	9	1101716330

Table 1: Space size v. power

We introduce additional parameter, s to denote the number of trust links that the attacker can access, instead of full control assumption. This reduces the space, such that:

$$|\mathcal{X}| = \binom{\rho - 1}{s - 1} \binom{|O^{0:t}_{tr\rightarrow .}| \cdot k}{s} \tag{7}$$

First term is the strict compositions of ρ into s parts for counting number of ways to divide the number of observations p into s parts. This number is then multiplied with number of ways this divided attack can be applied to the evidence that principal has. Even though, having this second parameter reduces the space, still exploring this space in a brute-force manner is not feasible. We consider exploring this spaceas an optimisation problem, which is strongly non-convex (and in many cases, discrete rather than continuous e.g., when the scores are discrete values), and thus, we cannot rely on standard gradient-based solutions either. Hence we turn to the usage of sampling based techniques. Among these, bandit optimisation is one of the most promising techniques for this problem.

To address this, we adapt a hierarchical optimistic bandit optimization HOO (cf., Algorithm 1) by Bubeck et al. [BMSS11] which provides lower regret bounds with quadratic computational complexity to strategically sample this space to estimate the mean-payoff gain of the attack in the attack space $\mathcal{X}$.

The main idea of this strategy is to estimate the gains of attacks accurately in the maxima, the partition that we are interested on, and loosely in the remaining partitions of the attack space. To achieve this, a binary tree which represents the partitions of the attack space $\mathcal{X}$ is created by the optimization strategy (as shown in Algorithm 1). Deeper nodes of the tree represent a smaller subset of attacks in the space. Each round a new node is added with some statistical information, which is updated each round. These statistical information is used for selecting a new region to explore more. Thus, the space is explored incrementally within the most rewarding sub-regions.

The nodes in the binary tree are indexed with integer pairs (h, i), where the depth of a node is h and i is to denote the index of all possible nodes at the depth (in range $1 \leq i \leq 2^h$). Therefore, the root node is represented as (0,1). The children of the root node are denoted as $(1, 1)$ and $(1, 2)$. In general, children of (h, i) are the nodes $(h + 1, 2i - 1)$ and $(h + 1, 2i)$. Regions of $\mathcal{X}$ are associated with each node and shown as $\mathcal{P}_{h,i} \subset \mathcal{X}$ and must satisfy the constraint, $\mathcal{P}_{h,i} = \mathcal{P}_{h+1,2i-1} \cup \mathcal{P}_{h+1,2i}$ for all $h \geq 0$ and $1 \leq i \leq 2^h$.

The statistical information that is stored in the nodes are:

- $T_{h,i}$: number of times an attack from the region (h, i) and its descendants are explored.
- $U_{h,i}$: initial estimate of the region (h, i), which is sum of:
  - $\hat{\mu}_{h,i}$: average reward received in the region associated with (h, i)
  - $\sqrt{(2\ln n)/T_{h,i}}$: corresponds to the uncertainty of rewards of the average and
  - $\nu_1 \rho^h$: corresponds to maximum variation of mean-payoff function in the region $\mathcal{P}_{h,i}$
- $B_{h,i}$: actual estimate of the mean-payoff function, calculated by $U_{h,i}$.

In particular, with the stored statistical information, the algorithm is progressively exploring the region as follows: Each round, the region to explore is selected by picking the node with the highest B-value (6-10 lines of Algorithm 1). The region that is selected is played and the corresponding reward is received (15-17 lines). The tree is updated with the previous statistics given the collected reward (19-33 lines). In detail, the path (a set of nodes) followed to select the region is updated with the reward in lines 19-21. The tree structure of space $\mathcal{X}$, $\mathcal{T}$ is traversed and the initial estimate of subregions are updated in lines 23-25. Initial optimistic estimates for new descendants of the selected node (H, I) is set in lines 26-27. B-values are updated in lines 29-32.

Algorithm 1 The HOO Strategy

Parameters: Two real numbers $v_1 > 0$ and $\rho \in (0, 1)$, attack space $\mathcal{X}$, all evidence $\mathcal{E}$.

Auxiliary function LEAF($\mathcal{T}$): outputs a leaf of $\mathcal{T}$

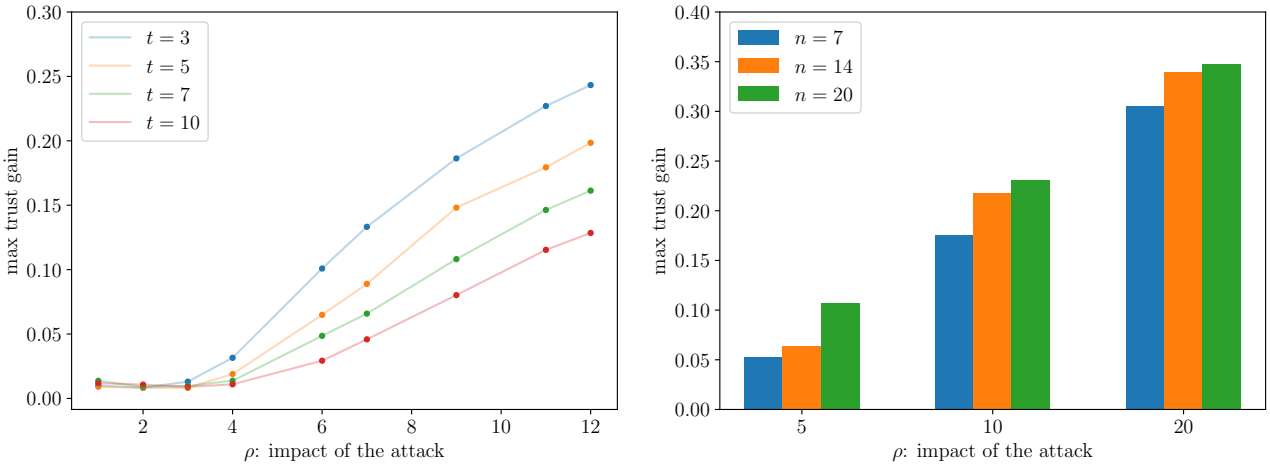
Initialisation: $\mathcal{T} = \{(0, 1)\}$ and $B_{1,2} = B_{2,2} = +\infty$.

for $n = 1, 2, ...$ **do** ▷ For each round n
  $(h, i) \leftarrow (0, 1)$
  $P \leftarrow \{(h, i)\}$
  while $(h, i) \in \mathcal{T}$ **do**
    if $B_{h+1,2i-1} > B_{h+1,2i}$ **then**
      $(h, i) \leftarrow (h + 1, 2i - 1)$
    else if $B_{h+1,2i-1} < B_{h+1,2i}$ **then**
      $(h, i) \leftarrow (h + 1, 2i)$
    else
      $(h, i) \leftarrow$ choose a child randomly
    end if
    $P \leftarrow P \cup \{(h, i)\}$
  end while
  $(H, I) \leftarrow (h, i)$
  Arbitrary choose one of corresponding attacks $\mathcal{E}'$ in space $\mathcal{X}$ with respected to partition (H, I)
  $Y = \mathbb{E}[\theta_{tr\to te}|\mathcal{E}'] - \mathbb{E}[\theta_{tr\to te}|\mathcal{E}]$ ▷ Receive corresponding reward for selected attack
  $\mathcal{T} \leftarrow \mathcal{T} \cup \{(H, I)\}$
  for $(h, i) \in P$ **do**
    $T_{h,i} \leftarrow T_{h,i} + 1$
    $\hat{\mu}_{h,i} \leftarrow (1 - \frac{1}{T_{h,i}})\hat{\mu}_{h,i} + \frac{Y}{T_{h,i}}$ ▷ Update the mean $\hat{\mu}_{h,i}$ of node (h, i)
  end for
  for $(h, i) \in \mathcal{T}$ **do**
    $U_{h,i} \leftarrow \hat{\mu}_{h,i} + \sqrt{(2\ln n)/T_{h,i}} + \nu_1\rho^h$
  end for
  $B_{H+1,2I-1} \leftarrow +\infty$
  $B_{H+1,2I} \leftarrow +\infty$
  $\mathcal{T}' \leftarrow \mathcal{T}$
  while $\mathcal{T}' \neq \{(0, 1)\}$ **do**
    $(h, i) \leftarrow$ LEAF($\mathcal{T}'$)
    $B_{h,i} \leftarrow \min\Big\{U_{h,i}, \max\{B_{h+1,2i-1}, B_{h+1,2i}\}\Big\}$
    $\mathcal{T}' \leftarrow \mathcal{T}' \setminus \{(h, i)\}$
  end while
end for

4 Simulation Results

To evaluate the performance of our attack model, we conducted empirical experiments to investigate our attack model in terms of: how prior evidence and the number of trustees in the environment affect the impact of an attack and how our attack model performs before converging to the best possible attack. Each experiment is repeated 1000 times with different trustee reputation parameters $\theta_{tr \to te}$ and average was taken to minimise the influence of random trustee parameters. All experiments included a simulated environment where an attacker wants to influence a truster's trust assessment of a single trustee. The targeted truster is making trust assessments with HABIT model, as described in Section 2 and the attacker is modifying the truster's observations of the environment by using our attack model, as described in Section 3.

In each experiment, the number of rounds was allocated to HOO Strategy was equal to the number of possible attacks. We set a high number of rounds to ensure convergence of the HOO Strategy, which can be seen that convergence was achieved as a matter of fact in a smaller number of rounds (as shown in Figure 2). The truster only had prior information about each trustee from the previous direct observations. Due to computational time available for experiments, s, the number of trust links that were affected by the attack, was picked as 2.



(a) Prior evidence, $n = 20$

(b) Number of trustees ($|O^{0:t}_{tr \to .}| = 2$)

Figure 1: Trust gains by attacks in varying dimensions:

- the number of trustees: n
- and the number of total previous evidence: t

4.1 Prior evidence

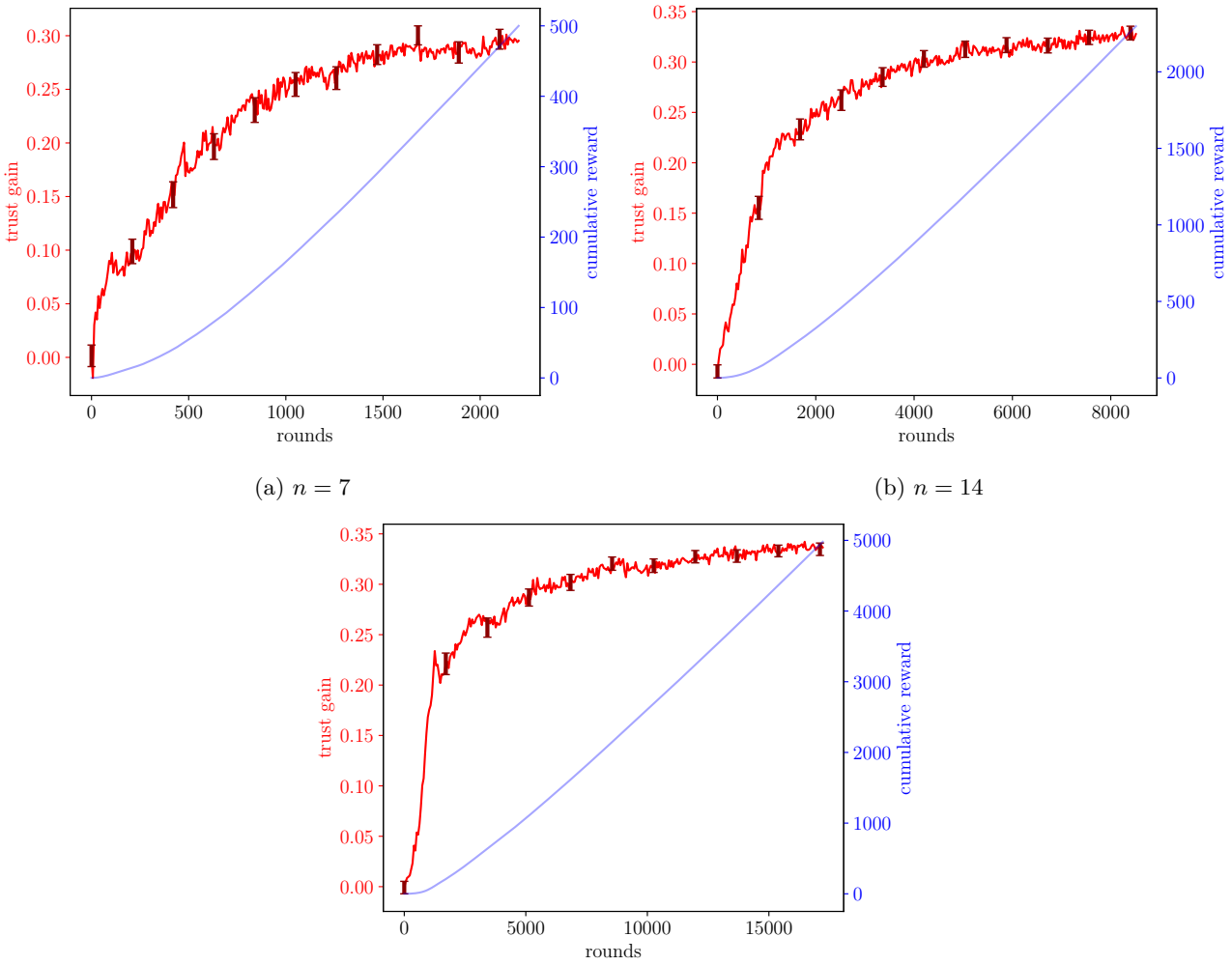
To determine the effect of prior evidence in the environment to our attack strategy, we ran a series of experiments in which prior information that truster had about trustees were varied in values of 3, 5, 7 and 10. The impact of the attack, ρ, were ranging from 1 to 12. As shown in Figure 1a, in the case of the number of prior direct interaction is being higher than the impact parameter $|O^{0:t}_{tr \to .}| > \rho$, the impact of attacks remained less than 0.10. The increasing trust gains were achieved when the attack budget overwhelmed the prior information available.

4.2 Population in the environment

HABIT trust model takes advantage from the similarities in the behaviours of trustees. Their results show that the number of trustees when correlation exists (similar reputation parameters, $\theta_{. \to te}$), model achieves more accurate trust assessments This advantage of the model turns out to be a disadvantage in an orchestrated attack which our attack model collected higher trust gains with the generated attacks in more crowded trustee populations (as shown in Figure 1b).

4.3 Performance over time

We investigated the performance of the attack strategy in each round and observed to the rewards of the strategy in the intermediate steps for the search for an optimal attack. In each of the varying number of trustees, HOO strategy behaved similarly in terms of trust gain and cumulative reward. As the space grew when the number of trustees increased and the number of rounds required likewise grew proportionally. It can be observed from Figure 2, a set of eligible attacks were explored throughout rounds. These can be used at any point depending on the resource of the attacker.



(a) $n = 7$

(b) $n = 14$

(c) $n = 20$

Figure 2: Trust gains of HOO Strategy in each rounds with simulation settings: varying number of trustees n and $\rho = 20$.
Error bars show standard error of the mean trust gain

5 Discussion

In our investigation, we considered a trust problem where an observer (truster) is making inference of risk of interacting with service providers (trustees). The orchestrated attack, consisted the following assumptions:

Total Visibility: We assume that the attacker has access to observe whole environment, but only had the ability to modify a subset of the information in the environment. On the other hand, many trust and reputation systems tend to publish such historical knowledge in practice. In addition, the trust model that

we are employing requires this assumption to make assessments. Therefore, for this empirical experiments, our environment were fully observable. This may not be the case in some trust and reputation systems.

Static Environment: As in many trust systems, the dimension of time in interactions is meaningful. This dimension can be exploited with on-off attacks, in which agents delay the interactions, meanwhile profiting by malicious behaviour. After a malicious period, for instance, service providers may leave the system and create a new identity to *whitewash* to previous poor reputation [WMLZ14] is not considered. One direction for future research can be investigating these types of attacks and devising an approach to detect and strengthen models of trust.

Knowledge about Trust Model: We assume that either the attacker has already known the trust model that is employed by the system or the attacker creates a virtual equivalent of the scenario in the targeted trust system. It is a strong assumption, however we would like to investigate further when such information is not available to the attacker. From the attackers perspective, question of how to predict the trust model that is used by the trust system and from moderators of the trust system perspective, question of how to devise a trust system to hide its predictability is intriguing and it is in our future research agenda.

Heuristics to Reduce Attack Space: Heuristics can be employed to reduce the attack space, thus yielding a higher performance in our attack model. For instance, as in our case, knowing that the targeted system is using HABIT model, an attacker may incorporate a similarity metric and create a smaller attack space, where the convergence can be achieved with fewer rounds. It is a realistic addition to our attack model that we would like to explore.

It is important to note that the attack strategy that we devised is compatible with other computational trust models. Other areas for future research are: investigating other trust models, trust-aware decision processes (e.g. such as [GNTT17, SRR15]) and verifying its effectiveness of our attack strategy with more realistic assumptions and creating defensive mechanisms to such sophisticated attacks.

6 Conclusion

In this paper, we have introduced a new evaluation technique with a generic attack model to model of trust. We investigated the performance of our adaptation of HOO strategy as an attack model for finding optimal attacks for a selected trust model. Our experiments showed that our attack model can exploit the strengths of the employed trust model. In addition, this is the first contribution to show the potential of devising a more realistic benchmark for models of trust. In future research, we will address the assumptions that were discussed and at the same time improve the performance of our attack strategy.

References

[BL16] Amir Jalaly Bidgoly and Behrouz Tork Ladani. Modeling and quantitative verification of trust systems against malicious attackers. *The Computer Journal*, 59(7):1005–1027, 2016.

[BMSS11] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *J. Mach. Learn. Res.*, 12:1655–1695, July 2011.

[BNS10] Chris Burnett, Timothy J. Norman, and Katia Sycara. Bootstrapping trust evaluations through stereotypes. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, pages 241–248, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems.

[FBZ13] Hui Fang, Yang Bao, and Jie Zhang. Misleading opinions provided by advisors: Dishonesty or subjectivity. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 1983–1989. AAAI Press, 2013.

[Fer73] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[GNTT17] Taha D. Güneş, Timothy J. Norman, and Long Tran-Thanh. Budget limited trust-aware decision making. In Gita Sukthankar and Juan A. Rodriguez-Aguilar, editors, *Autonomous Agents and Multiagent Systems*, pages 101–110, Cham, 2017. Springer International Publishing.

[HM04] Silvia Heubach and Toufik Mansour. Compositions of n with parts in a set. *Congressus Numerantium*, 168:127, 2004.

[HZNR09] Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. A survey of attack and defense techniques for reputation systems. *ACM Comput. Surv.*, 42(1):1:1–1:31, December 2009.

[JG09] Audun Jøsang and Jennifer Golbeck. Challenges for robust trust and reputation systems. In *Proceedings of the 5th International Workshop on Security and Trust Management (SMT 2009), Saint Malo, France*, page 52. Citeseer, 2009.

[Jøs12] Audun Jøsang. Robustness of trust and reputation systems: Does it matter? In Theo Dimitrakos, Rajat Moona, Dhiren Patel, and D. Harrison McKnight, editors, *Trust Management VI*, pages 253–262, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[KC09] Reid Kerr and Robin Cohen. Smart cheaters do prosper: Defeating trust and reputation systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, pages 993–1000, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.

[KSGM03] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 640–651, New York, NY, USA, 2003. ACM.

[MWLZ16] Tim Muller, Dongxia Wang, Yang Liu, and Jie Zhang. How to use information theory to mitigate unfair rating attacks. In Sheikh Mahbub Habib, Julita Vassileva, Sjouke Mauw, and Max Mühlhäuser, editors, *Trust Management X*, pages 17–32, Cham, 2016. Springer International Publishing.

[RD16] Yefeng Ruan and Arjan Durresi. A survey of trust management systems for online social communities - trust modeling, trust inference and attacks. *Know.-Based Syst.*, 106(C):150–163, August 2016.

[RPC06] Kevin Regan, Pascal Poupart, and Robin Cohen. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1206–1212. AAAI Press, 2006.

[SRR15] Sandip Sen, Anton Ridgway, and Michael Ripley. Adaptive budgeted bandit algorithms for trust development in a supply-chain. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 137–144, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems.

[TLRJ12] W.T. Luke Teacy, Michael Luck, Alex Rogers, and Nicholas R. Jennings. An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling. *Artif. Intell.*, 193:149–185, December 2012.

[WMLZ14] Dongxia Wang, Tim Muller, Yang Liu, and Jie Zhang. Towards robust and effective trust management for security: A survey. In *Proceedings of the 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, TRUSTCOM '14, pages 511–518, Washington, DC, USA, 2014. IEEE Computer Society.

[WMZL15] Dongxia Wang, Tim Muller, Jie Zhang, and Yang Liu. Quantifying robustness of trust systems against collusive unfair rating attacks using information theory. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 111–117. AAAI Press, 2015.