Exploring the Role of Witness Interactions in Heterogeneous Trust-Aware Societies

Botti Alessandro,[1] Leoni Eugenio,[2] Lorenzetti Simone,[3] Puschiasis Priamo[4]

[1] *Bocconi University, id: 3115589*
[2] *Bocconi University, id: 3119306*
[3] *Bocconi University, id: 3118280*
[4] *Bocconi University, id: 3118686*

Abstract. Trust is multifarious concept which plays a key role in the developing of a single human being as well as of the society as a whole. It is therefore relevant to study and develop a better understanding of it. We aim to do so by building on previous literature, implementing a model characterized by heterogeneous agents, with multiple decision processes. Previous studies have shown how giving more than one dimension to the concept of trust can affect a society's structure. We further study this idea, showing that agents can use additional dimensions of trust to directly improve their utility, by taking advantage of the experience of other trustworthy agents in order to avoid interactions with ill-intentioned ones

***Declaration**:*
In submitting this assignment:
1. We declare that this written assignment is our own work and does not include (i) material from published sources used without proper acknowledgment or (ii) material copied from the work of other students.
2. We declare that this assignment has not been submitted for assessment in any other course at any university.
3. We have a photocopy and electronic version of this assignment in our possession.

1 Introduction

Trust is an essential element of every society; on the individual level it is a critical factor in any relationship in which the trustor does not have direct control over the actions of a trustee and the environment is uncertain [1]. A wealth of literature has emerged on the topic, contributing with models where agents make use of trust and reputation systems in deciding how, when and who to interact with in a specific context [2]. The main goal of these models is to detect malicious agents in order to avoid interacting with them [3]. Some models try to do so by implementing a reputation mechanism, where reputation can be defined as the opinion or view of someone about something [4]. To our knowledge, though, only some of these works have studied how a reputation system affects the aforementioned goal when reputations aren't always given in a cooperative way [5] [6].

We will explore the co-existence of two types of trust-values, one for direct interactions and one for the reliability of reputation values, operating in the framework developed by Abari and White [7][1]. The main contribution of the two authors is the introduction of naive agents. Indeed, up to that point, the models developed mostly took into consideration two rational types of agents: trust-aware and ill-intentioned [8]. However, it is reasonable to assume that some members of society are not actually able to interact properly with other agents, notably: "naive agents are naive in terms of deciding how, when and who to interact with while always cooperating with other agents". Due to their nature, they aren't reliable when attributing reputation scores.

Given that their role in a system hasn't been fully analysed, we decided to build on the model proposed by the authors, who implemented a trust-variable measuring the reliability of agents, which we will use to extend the model to let the agents make use of a reputation system. This will allow us to investigate our research question, which is studying the effect of different types of reputation systems in a society with several types of agents, who aren't always reliable when providing reputation scores.

The paper is structured as follows. In Section 2 we analyze the relevant works in the literature on the topic, in Section 3 we describe the model we replicate and our extensions, in Section 4 we show which fixed parameter values we used, while in Section 5 we will report the results of the experiments.

2 Related work

Most trust models are built on the assumption that there exists only two type of agents: trust-aware and malicious. However, Abari and White [7] decided to go beyond this assumption and included a third type: the naive agent. In their work they analyzed the effects of naive agents on trust-aware individuals and the whole of society. They found out that malicious agents can be successfully isolated when naive agents are absent but when the proportion of naive agents exceed a threshold, malicious agents have the best utility in the society, consequently making defection the optimal strategy. They also show a mechanism through which trust-aware individuals can identify naive ones, notably, due to them always providing good reputation scores towards other agents. In our opinion, though, they fail to practically show whether and how this knowledge can be used by trust-aware agents, which is what we plan to add with the current work.

Yu and Singh [8] build a reputation model in which the problem of lack of information is dealt with through the use of Dempster's rule. Agents may keep two trust variables, one about reputations and the other about trustworthiness in direct interactions, but only the last one is used, if available. Furthermore, it is assumed that all agents provide truthful recommendations.

[1] For consistency, also the formatting of this paper is inspired by the same source

Huynh et al [9] investigated the role of trust in open multi-agent systems (MAS) in which agents are allowed to enter and leave the system at any time. In order to deal with the possible changes in the environment, they developed FIRE, a model which integrates different sources of information such as interaction trust, role-based trust, witness reputation, and certified reputation, to provide an extensive trust metric. However, FIRE as well is based on the stringent assumption that the agents in the model only exchange truthful information among themselves.

Some works relax this assumption, e.g. Yu et al. [10] and Sen and Sajja [11]. In the former, a trust model is formulated in large-scale peer-to-peer systems to detect malicious or unreliable peers. Interestingly, the authors not only explored different rating aggregation methods such as simple and exponential averaging, but they also considered how to effectively aggregate noisy ratings from independent or collusive peers using weighted majority techniques.

Sen and Sajja [11], meanwhile, face the problem of user agents selecting the right service providers to process tasks. They use a probabilistic approach which makes use of the summary statistics of the population distribution. Furthermore, each interaction is broadcasted to some other agents, with some noise on the actual performance. While we implement a somewhat similar idea for letting agents observe other interactions, we let these observations happen at random, with a given probability proportional to the agent's *Willingness to Observe*. Furthermore, the noisiness in our model will be 'intrinsic', in the sense that an agent tending towards cooperation can be observed when defecting, due to him interacting with a malicious agent.

In general, compared to the described works, we let agents be unreliable, so we use a trust variable for direct interactions and one measuring reliability. Additionally, we let agents address the lack of information problem through their heterogeneities, meaning that some will tend to trust unknown agents, while others not. We will also experiment with the possibility of agents observing other interactions, thus combining information coming from this process with the reputation given by other agents, instead of using only one of these. Furthermore, in some works the rating scores aren't actually used by the agents, as we plan to do to study the society structure that will emerge and the utility of different agent types.

As we have seen, some works, as the two most recently mentioned, are able to relax the assumption of agents always providing truthful reputation scores, doing so with models more complex than ours. However, we study the effect of reputation systems in a very different setting from theirs, as will become clearer when we will describe our model. The simplicity and generality of the model from which we build up can be an advantage, as it allows to better capture the implications of the experimented agent policies.

3 The Model

Given the nature of the research question we decided to implement an ABM. Indeed, it is the most used model in literature, as it allows to implement heterogeneities and study the effect of individual decision rules on agent groups as well as on the society.

We start our work by replicating the model in [7], which we hereby describe in a simplified manner for expositive convenience. It is to be noted that such model is used for scopes of theoretical investigations, thus not all the variables are interpretable in real-life terms. For example, while the purpose of a time-step is clear, its specific meaning can be derived only when applying the model to a field of interest, adapting it to reflect the characteristics of such scenario.

3.1 Environment Model

Connections. Each agent interacts with a subset of all agents given by his neighbourhood, which is a dynamic set with a maximum cardinality. At every step, each agent will send

connection requests to other random unvisited agents until the set cardinality is reached. Even when an agent has a full neighbourhood, he will still accept one incoming connection request at each step, using the FIFO criterion. Some agents will also end a relationship with another agent, removing the latter from the neighbourhood; we will explore this possibility when introducing agents and their methods.

Interactions. Two types of interchange can occur between couples of agents: Direct Interactions (DI) and Witness Interactions (WI). DI are "the most frequently used source of information for trust and reputation models" (ibidem, p.4), and have their own interpretation depending on the field of interest, e.g. they can be the buying and selling of a product in e-commerce. WI, instead, are used to let some agents ask and receive from their neighbours an assessment on the trustworthiness of a specific agent. Actually, the asking agent isn't interested in his neighbours opinion per se, he uses this interaction to assess their trustworthiness, in a dimension different from the DI one. Indeed, he asks them a rating on an agent he knows to be untrustworthy in DI, and checks whether or not his neighbours provide him with truthful answers.

Interactions are modelled using extensions of the Prisoner's Dilemma, a simultaneos non-zero-sum game in which each agent can either "cooperate"or "defect". WI take the form of the Generalized Prisoner's Dilemma (GPD), and don't provide any extrinsic reward. Instead, DI take the form of an Iterated Prisoner's Dilemma (IPD), and agents will earn a payoff based on Table 1.

Agent A / Agent B	**Cooperate**	**Defect**
Cooperate	**3, 3**	**0, 5**
Defect	**5, 0**	**1, 1**

Table 1: Payoff Matrix

Trust variables. As there are two different types of interactions, there also two distinct trust variables. We focus our explanations on $DIT_i(j)^t$, the trust value assigned by agent *i* to agent *j* at time *t* as a result of DI, but the same considerations apply to $WIT_i(j)^t$, used in the WI dimension.

$DIT_i(j)^t$, has the following properties: $1 \leq DIT_i(j)^t \leq 1$ and $DIT_i(j)^0 = 0$. This value has to be evaluated together with some thresholds, notably Ω_i and ω_i, where $-1 \leq \Omega_i < \omega_i \leq 1$. If $DIT_i(j)^t > \omega_i$, the agent *j* is considered *Trustworthy* by the agent *i* in DI, whereas, if If $DIT_i(j)^t < \Omega_i$, such agent is considered *Untrustworthy*. For any value in the middle-ground, the agent is considered *Not Yet Known.*

Following a DI, $DIT_i(j)^{t+1}$ will be the result of the updating scheme illustrated in Table 2.

Updating Equation	***Rating***	***Behavior***
$DIT_i(j)^t + \alpha_D(i) * (1 - DIT_i(j)^t)$	$DIT_i(j)^t > 0$,	*cooperation*
$(DIT_i(j)^t + \alpha_D(i)) / (1 - min(\|DIT_i(j)^t\|, \|\alpha_D(i)\|)$	$DIT_i(j)^t < 0$,	*cooperation*
$(DIT_i(j)^t + \beta_D(i)) / (1 - min(\|DIT_i(j)^t \| , \|\beta_D(i)\|)$	$DIT_i(j)^t > 0$,	*defection*
$DITi,j(t) + \beta_D(i) * (1 + DIT_i(j)^t)$	$DIT_i(j)^t < 0$,	*defection*

Table 2: Updating equation

In the formula, α_D is a positive weighting coefficient for cooperation in DI and β_D is a negative weighting coefficient for defection in DI. Following [3], we require that $|\alpha| < |\beta|$, under the idea that it is easier to lose trust than it is to gain it.

3.2 Agent Model

We have framed the backbone of our model, it is now needed to provide details on the agents which will populate it.

Agents. Four agent classes are defined: Trust-Aware (TA), Trust-Aware+ (TA+), Malicious (MA) and Naive (NA). Malicious agents always try to fool the other agent and get the highest reward. They will try to exploit as many agents as possible, thus they their set cardinality is the highest, equal to 100. Furthermore, due to their greedy nature, they won't drop any connection. Naive agents as well, because of their unsophisticated nature, are not able to end a negative relation with a ill-intended agent once it is established. They have an intermediate neighbourhood set cardinality of 15. Trust-Aware agents act following more complex decision rules, and they will drop agents *Untrustworthy* in DI. Being more selective, they have the lowest set cardinality, equal to 5. Finally, Trust-Aware+ are identical to TA in all aspects, except that they also drop agents *Untrustworthy* in WI.

Direct Interaction Methods. The decision rule followed by NA and MA agents in DI is straightforward: the former agents will always cooperate, while the latter will always defect. As for TA and TA+, they will adapt their behaviour based on the trustworthiness of their counterpart. If the interacting agent is *Not Yet Known*, the strategy is to copy his last move if there is one, otherwise to start with a cooperation. If the other agent is deemed *Untrustworthy* in DI, then TA and TA+ will defect, while the will cooperate if they deem him *Trustworthy* in DI.

Witness Interaction Methods. When it comes to WI, inquiring agents will ask neighbours to rate an *Untrustworthy* agent, while interrogated neighbours have to decide whether or not to provide a rating close to their actual opinion. The authors identify three policies: Honest (Ho), Liar (Li) and Simpleton (Si). Referring to Algorithm 1, what changes is the value with which to replace the asterisk, respectively: "$DIT_i(j)^t$", "1 - $DIT_i(j)^t$", *and "1"*. TA and TA+ will be Honest, MA will be Liars, and NA will be Simpletons, meaning that TA and TA+ will always cooperate, by sending the actual DI trust value, MA will always defect, rating low agents who are *Trustwhorthy* and rating high other MA, while NA, by always rating high, will sometimes defect (when reccomending MA) and other times cooperate (by rating high other NA, TA or TA+). DT is set to 0.25 by the authors, meaning that a defection happens only if the answer is significantly different from the actual opinion of the neighbour.

Algorithm 1 Answering Policy[2]

if receiving a witness request about *j* from *k* **then**
  opinion = ∗
  send opinion to *k*
  if $|opinion - DIT_i(j)^t| < DT$ **then**
    Send cooperation
  else
    Send defection
  end if
end if

[2] This pseudo-code is from [4], p.7

3.3 Extending the Model

As we have seen, WI have the purpose of letting agents update their $WIT_i(j)^t$ scores, and the authors implemented them to show that TA+ manage to correctly identify NA agents, dropping them as they are *Untrustworthy* in WI. We plan to extend the role of these interactions for 2 reasons. Firstly, agents don't really have an incentive for cutting relations with NA, as they always cooperate in DI. Secondly, the authors show that trust-aware agents are generally able to reach an equilibrium without connections with MA, but this requires facing each one of them in DI until they are finally deemed *Untrustworthy*.

Thus, we define the "Thoughtful" (Th) characteristic, which makes an agent avoid accepting just any connection request. Instead, such agents will ask neighbours for a rating of the agent who required the connection. This rating follows the rules of WI interactions, so values are sent according to the neighbour policy and DI trust value towards that agent. The request will be accepted only if the average opinion surpasses ω_i, meaning that the "Th" agent deems him *Trustworthy* based on his reputation. If the neighbours opinion is perfectly mixed or if it falls in the category of *Not Yet Known*, the inquiring agent will append the examined agent to the end of the list with agents who asked for a connection, thus checking for him again in a while. Instead, inquiring agents will never accept an agent who, at any given step, has been deemed *Untrustworthy*. We also define the "Th+" characteristic (Enhanced Thoughtfulness), which works as just described, but will let agents take into account only the opinion of neighbours *Trustworthy* in WI.

Finally, we define the "Observer" trait ("Ob"), which allows agents to observe others interacting, forming a prior modelled, for convenience, as $DIT_i(j)^t$. Notably, each agent with this trait will have a static attribute named *Willingness to Observe*, defining the probability to observe an interaction, for all interactions that play-out at a time step. Furthermore, such agents will not just randomly send connection requests; instead, they will randomly select an agent and send a connection request to him only if he has a reputation score above the *Trustworthy* threshold. Reputation is calculated as the sum of two factors, a prior coming from observed interactions involving that agent, plus the average of the trust values returned by the neighbours[3].

More formally, defining the asking agent as *i*, the reputation score available to *i* about agent *j* as $R_i(j)$, the opinion value of neighbour *k* towards agent *j* as $S_k(j)$, the prior as p_i, the neighbourhood of *i* as N and its length as n, he will send the request if $R_i(j) > \omega_i$, where $R_i(j) = p_i + \frac{1}{n}\sum_{k \in N} S_k(j)$.

4 Parametrization

We use the same experimental values as in [7], to operate in the same testbed. Most of the values are provided, as we described in the previous section for convenience. Meanwhile α, β and values for the omegas are not expressely specified, so we looked to [3] and [12], to which the authors refer. By doing so we come out with a set of values for α and β, which we calibrated on the experiments ran by the authors. These parameters define the speed at which agents classify each other into being *Untrustworthy* or *Trustworthy*, but we find that our model converges faster than theirs despite the particular values used. While this is unfortunate, we don't deem it too big of an issue, as the overall dynamics and trends are well replicated. Thus we use $\alpha = 0.2,\ \beta = -0.3$ as they are the ones more in line with [3], and we will perform sensitivity on the other combinations.

[3]If the agent with the "Observer" trait has also the "Th+" trait, also in this case he will listen only to neighbours *Trustworthy* in WI

As for Ω_i and ω_i, we found no mentions on how to initialize them, so, while respecting $\Omega_i < \omega_i$, we calibrated them by testing different possibilities: letting them be distributed according to a uniform, to a gaussian, or letting them be homogeneous with a pre-speecified value. We discarded the first option because it gave results very stochastic and different from the author's ones. While the last option didn't perform poorly, we've opted for best in the class of the second option, notably a gaussian centered at 0.3 for ω_i and at -0.3 for Ω_i, with equal variances of 0.5. Indeed, this solution gives satisfactory results and appears more in line with the context of heterogeneity, and with definition of such thresholds [12]. Furthermore, it seems reasonable that most agents will have $\Omega_i < 0$ and $\omega_i > 0$, meaning that they consider an unkown agent as *Not Yet Known*. At the same time, this allows also for some agents to consider unknown others as *Trustworthy* or *Untrustworthy*; we define these agents as S*uper-Trusters* and *Skepticals*. This allows also for interesting behavours, for example two TA agents will generally endlessly cooperate, but, if both are *Skepticals*, then a single defection at the beginning can be enough to snowball into one dropping the other, missing out on a good relationship due to their conservative natures colliding.

We find relevant to mention here a counter-intuitive result obtained when calibrating on Experiment 3. Notably, while we agree with the authors interpretation saying that, if enough Naive agents populate a society, Malicious will have the highest average earnings, we argue that this is not necessarily true with the population percentages they use. Indeed, despite there being more NA than MA, and despite connection requests being sent at random to other unvisited agents, we notice that about 70% of MA agents will still fill their neighbourhoods mostly with other MA, thus hindering their average earnings. The reason for this comes from the fact that, in the long run, only MA will send connection requests. This is relevant because if an MA sends a connection request to an NA, then one MA will have one NA in his neighbourhood, while if the request is sent to one of his peers, then there will be two distinct MA with an MA in their neighbourhood. Thus, despite the latter situation being rarer than the former, it counts double in some sense. Furthermore, if an MA receives a request from another MA, he will need to send less connection requests to fill his neighbourhood, meaning that he will have less chances to equilibrate his neighbourhood composition.

5 Results

We ran 3 different experiments and performed sensitivity analysis on each one of them, investigating the effect of different methods and parameter values on several factors. Notably, we seek for effects on the number of MA encountered by trust-aware agents, on the network structure of the society as a whole and on the average utility by type of agent, which corresponds to the wealth earned per step by an agent normalized by his neighborhood and averaged across agent type. Indeed, we define the average utility of an agent as his wealth increase at a given step averaged across his interactions.

Experiment 1. We run a simulation which is similar to Experiment 1 in [7], with TA being 66% of the population and MA being 34%. Differently from the authors, here we let TA agents have the "Th" trait. The leading question for this Experiment was whether this new connection policy would help or deceive TA in isolating themselves from MA.

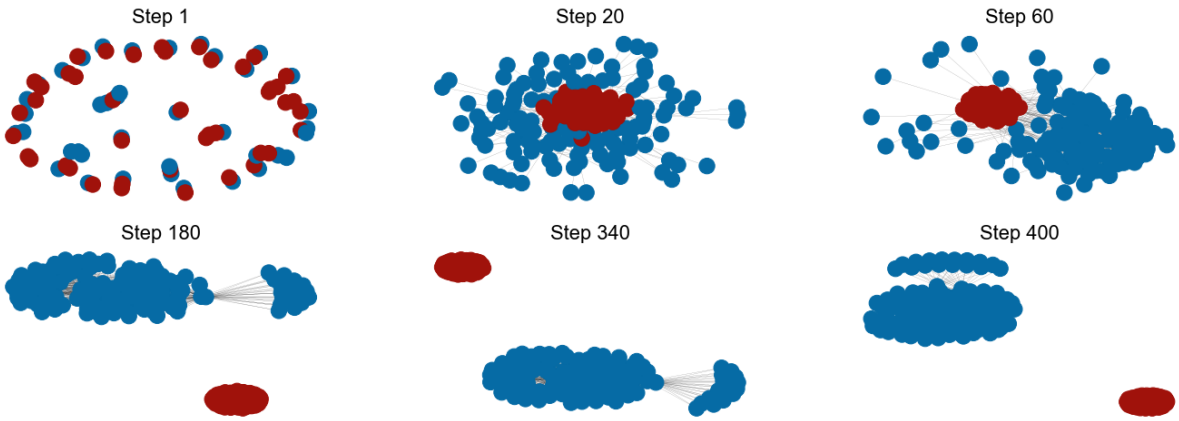


Figure 1: Network structure at several times steps, blue circles represent TA, while red ones are MA

As we can see from Figure 1, here TA agents isolate themselves much before step 400, indeed, we note that they need to interact with only with about half of the MA population. We can also see that a particular network structure arises: some agents appear to have many more neighbours than others, managing to connect and gradually integrate into the TA society an otherwise separated cluster of TA agents; more in general, they act as bridges. This behaviour is due to the heterogeneity in omega values, but we need to describe thouroghly what happens in order to understand the dynamics involved.

TA and MA send connection requests at random, but TA will not accept other agents unless their reputation puts them above the *Not Yet Known* status. Since, if there are no neighbours, the reputation score is set to zero, thus only S*uper-Trusters* TA will accept other TA in the beginning, and we found them to be around 8% of the TA population. These agents will keep accepting other agents, ending up with a neighbourhood up to 10 times bigger than the cardinality of 5. This happens because most TA will have difficulties filling up their neighbourhoods, as, aside from S*uper-Trusters*, only other MA will accept them, and will eventually get dropped. These accepted MA, until they are gotten rid of, will also try to lower the reputation scores of other TA while boosting that of other MA, though this requires that an interaction occurred between them, so it won't happen often in the early steps. Instead, when *Super-Trusters* manage to build enough connections, they will be able help TA bond with their peers with whom *Super-Trusters* have interacted with, by upvoting them. When this happens, TA will have some pending connection requests from their peers, since they postponed their decisions for all *Not Yet Known* agents, meaning that they will easily fill their neighbourhooods with other TA.

Meanwhile, *Skepticals* will drop all incoming connection requests until they are accepted by an agent who can convince them of the trustworthiness of another agent. Furthermore, they will drop another TA if the latter doesn't convince them fast enough of his trustworthiness. Some of these agents will not manage to fill up their neighbourhood by step 400, some of them may even end up alone. To avoid this, it is crucial that they manage to maintain a connection with another TA early enough, expecially if that TA is a S*uper-Truster*.

To stess the importance of S*uper-Trusters*, we ran sensitivity analysis on omega values, finding that, when each agent considers a newly found agent as *Not Yet Known*, not a single TA will manage to build even one relationship with another TA.

Experiment 2. We operate in the same framework of the previous one, but we add Naive agents to the model, with a population comprised of 34% TA, 33% MA and 33% NA. By reccomending any agent, they act as bridges, but, differently from S*uper-Trusters*, they can help MA agents to take over the neighbourhoods of TA ones. At the same time, they can speed up the process of TA filling-up their neighbours with other TA, helping them in closing against MA agents.

We find that, in this more complex society, some TA may never reach a stable equilibrium, as we can see in Figure 2. Indeed, if they have even one NA in their neighbourhoods, they won't manage to protect themselves from incoming MA requests. This is because other TA may have not interacted with such MA, so they won't be able to negatively rate them, while NA will always push TA into accepting other agents. Thus, Naive agents can compromise TA utility, almost nullifying the advantages of the "Thoughtful" trait.

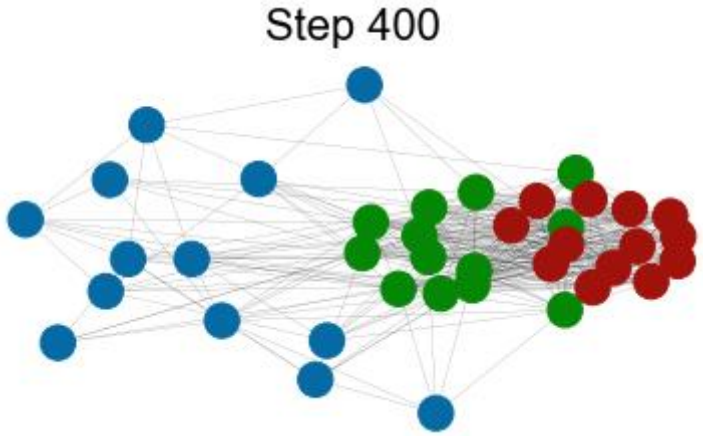
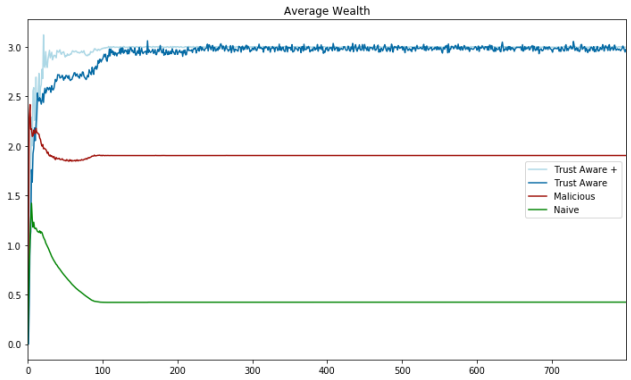


Figure 2: network state at step 400 for a population of 50 agents. TA in blue, NA in green and MA in red

Experiment 3. We run a simulation with 15% TA+ who have the "Th+" trait, 15% TA who have the "Th" trait, 50% MA and 20% NA. In this harsh environment, we test whether taking advantage of WI to listen only to the opinion of agents trustworthy in that dimension can help TA+ avoid the shortcomings shown in the previous experiment.

As we see through the average wealth increases, shown in Figure 3 on the left, TA+, despite having a stricter, thus slower, connection mechanism, manage to fill their neighbourhoods with agents trustworthy in DI even faster than TA do, thus earning the cooperation reward of 3. MA agents have their highest average earning during the very early steps, when they manage to get away with their defections. As TA and TA+ adapt their strategies, though, their utility reaches its minimum, before increasing again as they remain connected only to NA agents. TA are the only ones who, even in late time-steps, will have relevant spikiness in their utility. This is due to what we have seen in Experiment 2, indeed, some TA may get in contact with a MA agent at any time step.

TA+ are also able to avoid significantly more MA, on average, than TA with "Th" trait do, as shown in Figure 3, on the right. The interpretation is that, as long as they have some TA in their neighbourhood, they get fooled less often by NA. Indeed, a TA+ can get fooled into accepting an MA only if he is a *Super-Truster* and trusts the reputation given by a newly connected NA. This result isn't obvious, since the mechanic for accepting connection requests requires time to play out. Until then, TA+ neighbourhoods are filled thanks to the connection requests they send, and they have a good probability of sending it to MA agents in this scenario. Nevertheless, a simple check on incoming connection requests seems to be enough to significantly reduce self-exposure to malicious agents.



Type	MA_encoutered	Num_Agents
MA	78.320000	100.0
NA	57.650000	40.0
TA	67.666667	30.0
TAplus	21.600000	30.0

Figure 3. On the left is the average wealth increase by agent type, plotted for each time-step, on the right is the average number of malicious agents encountered by each type of agent

Experiment 4. Finally, we test, in the same setting, an even stricter connection rule. Indeed, now TA+ will have also the "Observer" trait and a *Willingness to Observe* equal to 0.001, making them more judicios when sending connection requests. This is a double-edged sword, as, by being so carfeful with connection requests, they may spend quite some time without any neighbour.

We find that the result of this addition are stochastic, and don't always reduce the average number of MA encountered. We speculate that this is because *Super-Trusters* will bypass these strict rules in the beginning, sending connection requests and accepting anyone.

Indeed, we find that if we force every agent to consider an unvisited agent as *Not Yet Known*, TA+ with "Observer" trait will manage to avoid any contact with MA agents! This happens because, from the beginning up until even time step 300, they will only connect with an agent after having seen enough of his interactions with other agents to be able to classify it into *Trustworthy* or not. In this case, the total avoidance of MA agents leads to a domination of the "Ob" trait on the "Th+" one. Indeed, in order for the WI trust values to be updated, an agent must encounter at least an MA, so he can test the reliability of his neighbours. If this never happens, he will never be able to consider any agent *Trustworthy* in WI, thus he will never listen to his neighbours opinion, and will act solely based on his prior. While this is a powerful result for TA+ agents, relying solely on observations can lead to a very slow filling up of the neighbourhood, expecially if *Willingness to Observe* is very low. Hence, such agents will avoid negative interactions, but will also have zero utility for a long time, creating a trade-off between these two aspects.

Sensitivity. Since ours is a theoretical experiment, which doesn't use real world data, running sensitivity analysis is critically important. For each experiment we ran several models, changing the values of alpha, beta and of the mean of the gaussian distribution from which omega values are generated. We found that stochasticity generally plays a bigger role than the parameter values, but results are pretty consistent despite both factors. The parameter with the biggest impact seems to be beta, which usually doesn't change the model behavour, except when it is set to -0.1. This is consistent with what shown in [3], as such a low value for beta can make it harder to update the trust value enough to surpass a threshold

Conclusions

While most trust-models achieve the isolation of untrustworthy agents, we find that the necessary mechanisms can be sub-optimal for the agents. We find that using connection policies based on reputation scores and on an additional trust-variable capturing an agent's reliability can lead to interesting results. Indeed, such simple policies can lead to undesirable patterns, such as not being able to achieve a stable isolation due to naive agents, while more complicated

ones may bring-in some stochasticity or may be worthwhile only in some scenarios. At the same time, we find that there are some policies that allow agents to have a greater utility at each step, by letting them avoid contacts with malicious counterparts, achieving isolation without the need for a high number of Direct Interactions.

We also find that, when adding restrictions to the connection-policies, the role of *Super-Trusters* can become extremely impactful, as they can single-handedly determine the final shape of the society, acting as reliable bridges among heterogeneous agents.

It would be interesting for future research to build on this model, for example analysing it in an open system, where new agents can enter, checking if Zacharia and Maes [13] requirements are met. Furthermore, it would be interesting to evaluate the model with sneakier malicious agents and, in general, adding complexity to the decisions rules and updating equations, for example making it harder to lose trust once an agent has surpassed a certain trust threshold.

Bibliography

[1] J. D. F. S. R.J Mayer, “An integrative model of organizational trust,” *Academy of management review* , pp. 709-734, 1995.

[2] S. H. D. J. Ramchurn, “Trust in multi-agent systems,” 2004.

[3] B. S. M. Yu, “A social mechanism of reputation management in electronic communities,” *Klush M., Kerschberg, L. (eds) CIA 2000. LNCS (LNAI), vol 1860,* pp. 154-165, 2000.

[4] J. &. S. C. Sabater, “REGRET: a reputation model for gregarious societies,” *Castelfranchi, C. & Johnson, L. (eds.), Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems,* p. pp. 475–482, 2002.

[5] M. F. P. &. R. M. Schillo, “Using trust for detecting deceptive agents in artificial societies,” *Applied Artificial Intelligence, Special Issue on Trust, Deception, and Fraud in Agent Societies,* vol. 14(8), p. 825–848, 2000.

[6] S. B. A. &. D. S. Sen, “Believing others: pros and cons,” *Proceedings of the International Conference on Multi-Agent Systems,* p. 279–286, 2000.

[7] T. W. Amirali Salehi-Abari, “The Impact of Naive Agents in Heterogeneous Trust-Aware Societies,” 2009.

[8] B. &. S. M. P. I. Yu, “ "An evidential model of reputation management"," *Castelfranchi, C. & Johnson, L. (eds.), Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-AgentSystems,* vol. 1, p. 295–300, 2002.

[9] J. N. S. N. Huynh T.D, “An integrated trust and reputation model for open multi-agent systems,” *Autonomous Agents and Multi-Agent Systems 13(2),* pp. 119-154, 2006.

[10] S. M. S. K. Yu, “Developing trust in large-scale peer-to-peer systems,” *IEEE First Symposium on Multi-Agent Security and Survivability,* pp. 1-10, 2004.

[11] S. S. N. Sen, “Robustness of reputation-based trust: Boolean case,” *Castelfranchi, C. & Johnson,L. (eds.), Proceedings of the First International Joint Conference on Autonomous Agents and Multi-AgentSystems,* vol. 1, p. 288–293, 2002.

[12] P. S. Marsh, “Formalising Trust as a Computational Concept,” *PhD thesis,* vol. Department of Computing Science and Mathematics, no. University of Stirling, 1994.

[13] G. M. P. Zacharia, “Trust through reputation mechanisms,” *Applied Artificial Intelligence,* vol. 14, p. 881–907, 2000.