Community Enforcement and the Emergence of Trust

Carlos Rodríguez-Sickert*

April 13, 2004

Abstract

In this paper, the link between community enforcement and co-operation in environments where trust is required is investigated. For this purpose, an assymetric trust game in which peer-to-peer sanctioning mechanisms are available is considered. Firstly, the outcome of this game is characterised in terms of the structure of incentives and the trustees' behavioural dispositions to sanction deviant peers. Subsequently, within a payoff-dependant imitation framework, the interaction between trusters and trustees is modelled as a two-population recurrent asymmetric game where both a) the evolution of a social norm, which enforces trust honouring within the trustees' community and b) the emergence of trust between-communities are analysed as inter-dependant processes. Provided that the sanctioning mechanism is efficient enough, and that the initial group reputation is high enough, (honoured) trust will evolve and become a stable outcome.

JEL Classification: D64, D83, Z13.

Keywords: Trust, Group Reputation, Social Norms, Social Learning.

1 Introduction

Within the Rational Choice tradition, the emergence of trust in current economic life is called to be sustained by the long-term structure of incentives that these relations tend to embody. Furthermore, authors like Hardin ([10],[11]) will put trust and reputation as inextricably linked concepts. Hardin, concerned with the *sui generis* character of relations of trust, comes up with the idea of *encapsulated trust*. "In modal trust relationships, the trusted party has

*Assistant Professor at the Catholic University of Chile and International Fellow at the Santa Fe Institute, New Mexico. Direct correspondence to Universidad Católica de Chile, Instituto de Sociología, Vicuña Mackenna 4860, Macul, Santiago, Chile; Tel.: 56-2-6864657; e-mail: crodrigs@puc.cl. I thank Patha Dasgupta, Ugo Pagano, Simon Deakin, Christoph Kuzmics and specially Bob Rowthorn for their useful comments. This research was funded by the Cambridge Political Economy Trust Society and the Catholic University of Chile.

an incentive to be trustworthy, an incentive grounded in the value of maintaining the relationship into the future. That is, "my trust of you is encapsulated in your interest in fulfilling the trust." What differentiates trusting (as a belief) from mere expectations in his analysis, Hardin continues, is that "...my expectations are grounded in an understanding of your interests specifically with respect to me" [11, p. 3]. Thus, trust and individual reputation become inextricably linked as the reason for the trusted party to honour trust is grounded in the value of maintaining the relationship into the future[1]. The *sui generis* character of a Relation of Trust, following Baier [1], will be associated here with the truster's voluntary exposure to the trustee's moral dispositions that is required in these transactions to effectively take place. In this scheme, reputation and trust, although related, are not considered as inseparable concepts. Rather, reputation is understood as a particular case of an outcome-improvement device when group reputation is too low. Investment in individual reputation, as it is conveyed in the analysis of Dasgupta [5] and Kreps [16], can elicit trustworthy behaviour even in opportunistic agents in the existence of a group of intrinsically trustworthy individuals.

It is beyond question that the incentives associated with reputation investment in the achievement of social co-operation in exchange contexts where trust is required. Nonetheless, it is inaccurate to claim that only via reputation mechanisms social co-operation can be attained within these contexts. There is conclusive experimental evidence which shows that moral dispositions can operate as a countereffect of opportunism in trust environments (see, for example, Berg et al. [14] and Snijders and Keren [19]). In these environments, when individual reputation cannot be built, it is the reputation of the group that determines whether trust will emerge or not. Neither can be claimed that the investment in individual reputation solution applies to every relation of trust. In a significant number of economic situations the interaction is static or, if dynamic, acquires a recurrent form where agents from different groups cannot engage in repeated interaction to create an isolated transaction. Rather, the interaction takes place between the trusters and the trustees only as anonymous members of their respective communities.

To a certain extent, as argued in Dasgupta [5], in an anonimous interaction setting where agents cannot engage in repeated interaction, group reputation becomes a pure public good (no rivalry in its use and no exclusion is possible). Following this approach, one can go further and establish that the provision of the public good is successful when group reputation is high enough to induce trusters' participation in an eventual transaction. In this scheme, the specific problem that will be investigated in this paper is whether social pressure within the trustee's community might boost group reputation to the levels required for the emergence of trust when intrinsic group reputation is not high enough. By considering mutual punishment mechanisms within the trustees'

[1] The other case of encapsulation, for Hardin, arises when the truster has special reason to think that the trusted party might take her interests into account when making a decision –he mentions friendship and love as the feelings involved in this case. These scenarios should be homologated to Willliamson [21]'s concept of *personal trust*.

community, group reputation will become an endogenous variable. Whether trust will emerge or not will depend on the provision of this public good –group reputation–, which in turn will depend on the levels of enforcement –via ostracism for example– implemented within the trustees' community. In addition, by relaxing the common knowledge of moral dispositions, the issue of the possible divergence between effective group reputation and perceived group reputation is tackled.

The structure of the paper is as follows. In section (2), an archetypal case of a Relation of Trust is modelled as an asymmetric sequential game and its outcome characterised under the material self-interest paradigm. In line with experimental evidence, in section (3), the possibility that group reputation is enhanced via community enforcement is considered. Specifically, the evolution of a social norm which enforces trust honouring within the trustees' community and between-communities trust emergence are analysed as inter-dependent processes. First, a behavioural model is considered. Secondly, a social learning model in which agents imitate successful strategies is analysed. Finally, in section (4), the results obtained in the previous sections are put in context and its implications on a number of social settings, discussed.

2 Relations of Trust: An Analytical Framework

2.1 A relation of Trust as a Trust Game

In this article, a relation of trust is associated with a particular structure of interaction characterised by (i) the existence of potential *mutual benefits* associated with a co-operative outcome; (ii) *voluntary participation* on the side of the truster (relevant exit option); and (iii) *opportunistic incentives* on the side of the trustee. These three features, although resemble Coleman [4]'s standard definition, depart from his approach by requiring the existence of a relevant exit option on the truster's side is required. It is the voluntary character of the truster's exposure that triggers the specific moral obligation –of reciprocal nature– involved in a relation of trust.

This approach is put forward in formal terms by characterising the structure of incentives of a trust game where one trustee and one truster are involved in the transaction and the sequentially of play guarantees that placement of trust is observed before the trustee decides how to respond to this placement. Thus, we have that

Definition 1 *In a Trust game:*

(i) a transaction (or more than one) is available to two parties;

(ii) although, for a transaction to take place, it is required that the truster, agrees to participate (Place Trust), its final outcome is determined –after participation has been observed– by the trustee;

(iii) neither the strategy of avoiding any possible transaction (Distrust) dominates any form of placement of trust nor vice-versa;

(iv) in order to achieve a Pareto Superior outcome relative to the initial situation, the trustee has to incur in a cost (with respect to the option of dishonouring trust);

(v) we denote any action of the trustee that improves the situation of the truster at the expense of the trustee as Honouring of Trust, and any action that worsens it as Betrayal.

The first feature identified previously –the existence of *co-operative benefits*– is expressed in the definition above in the existence of a Pareto Superior outcome with respect to the no-transaction scenario. The second feature –the *voluntary* character of the trustee's participation– is expressed in the availability of a relevant exit option: distrust which is not strictly dominated by any form of trust placement (this is what gives the relevance attribute to such an option). Finally, participation involves *exposure* to the moral dispositions of the trustee, i.e., the existence of opportunistic incentives, not only misaligned with the trusters' interests, but with the quality that the exercise of this opportunism would result in a worse position for the truster with respect to the no-transaction scenario. This feature is expressed in the fact that the achievement of a Pareto Superior outcome –where both the trustee and the truster see their position improve– requires the trustee to incur a monetary cost[2].

In the simplest representation of a trust game, originally presented in Dasgupta [5], there are two players: the truster and trustee[3]. In this trust game, the truster decides whether or not to participate in a transaction with the trustee who decides whether or not he will honour trust if it is placed in him. What follows is a slightly modified version of Dasgupta's original representation.

Sequence of the game In the first stage, the truster decides whether or not to trust the trustee. In the second stage which is reached only if the truster places trust in the trustee, the trustee decides whether to honour this trust or betray it.

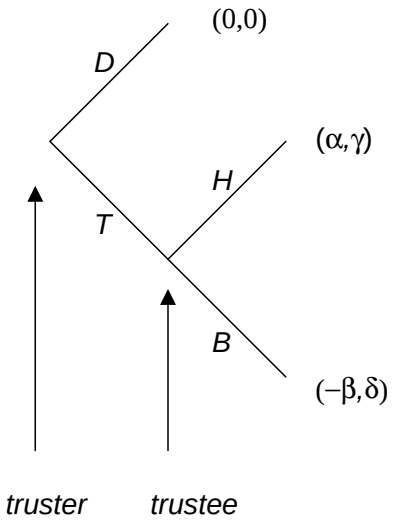
Payoff structure If a trustee decides to honour trust if trust is placed in him, associated with the monetary benefits of cooperation, he obtains a monetary payoff of γ; otherwise, if he betrays it, he obtains, associated with the monetary benefits of opportunism a payoff of $\delta > \gamma > 0$; the truster, if her trust is honoured enjoys a payoff of α; if her trust is betrayed she suffers a loss of $-\beta < 0 < \alpha$. If she does not participate in the transaction (distrust)

[2] In terms of the categorisation of trustees' actions, we follow Gouldner [9] understanding of a norm of reciprocity and, thus, even if it is not the most favourable action towards the trustee, any form of improvement is associated with honourment of trust as both principles of the norm are fulfilled. An action, which does not change the position of the truster is also associated with betrayal in the name of the second principle of the norm described by Gouldner.

[3] A similar representation is presented in Kreps [16]. In more recent literature, there is no definitive consensus regarding what a trust game is. However, many authors, e.g., Williamson [21], Snijders and Keren [19] and James [13], when referring to a trust game, do take this initial representation –or a simplified version– as the standard one.

her situation stays unchanged, i.e., she gets zero. When modelling a particular transaction we have to bear in mind that after placement of trust payoff parameters constitute the difference between the after transaction game and the non-transaction scenario. The categorisation of a relation of trust needs to be carried out after the normalisation of exit payoffs has been done as neglecting the difference in exit payoffs could be misleading. Consider for instance the decision of accepting/not accepting to do a job in exchange for an spot amount together with a delayed payment in the absence of a formal contract in two different settings: one in which the potential worker faces a monopsonist labor market versus one setting in which he faces a competitive market. Whilst in the latter, his participation in the transaction can be considered as a voluntarily one, in the former case this would not be the case[4]. In terms of parameter values, once exit options of the employees are set to zero, the existence of an alternative in the competitive setting could be expressed by a lower α.

In its graphic representation, the basic trust game acquires the following form:



The transaction is characterised by a structure of monetary incentives and sequentiality of play which has all the properties described in Definition (1):

1. Without placement of trust, no transaction takes place;
2. the condition $-\beta < 0 < \alpha$ implies that neither the truster's exit option (distrust) strictly dominates trusting nor trusting dominates distrust
3. the condition $\delta > \gamma > 0$ implies that the trustee faces opportunistic incentives as honouring trust becomes costly in monetary terms;

[4]Notice that for the example to be valid, we require that the on spot proportion of his salary gives her a higher payoff than the leisure option.

4. as α, β, γ and δ are positive, only the co-operative outcome (trust is placed, trust is honoured) is Pareto Superior to the no-transaction scenario.

Notice that the effect of the moral obligation on the trustee's payoffs has not been incorporated yet. It is the structure of the available actions that determines whether an interaction situation constitutes a Relation of Trust or not. As we assume that this effect will vary across the population of trustees, the setting will therefore change to one of incomplete information, i.e., the Basic Trust game (henceforth the BT game) will become a Multiple-type BT game.

Among the social settings, whose structures of monetary incentives can be represented by the BT game, we can find an important number of economic transactions.

Dasgupta [5], for instance, associates the roles of the truster and the trustee with a customer who decides whether or not to buy an used car and the salesman who decides whether to sell him a "lemon" or a "peach", qualities which are not observed by the buyer. In general, we could think of any transaction where there is an ex-ante asymmetry of information regarding the quality of the object of exchange. Most services which involve specialised human capital cannot be monitored (or monitoring is too expensive) and, thus, are characterised by this asymetry (e.g., the transaction between a patient and her doctor or between an expert economic consultant and the company that hires his services).

2.2 The Basic Trust game played between RATs:

The solution to the Basic Trust game provided by standard game theory is an unambiguous, but unpromising one. This prediction is expressed in the following proposition.

Proposition 2 *If the basic trust game is played between rational individuals driven by material self-interest (RATs) and this fact, the structure of the game and all the parameter values are common knowledge, then the unique sub-game perfect Nash equilibrium (SPNE) is (Distrust, Betray if Trust is placed).*

Proof. In the second stage of the game a self-interested agent will betray trust if it is placed in him as $\delta - \gamma > 0$. The truster –given common knowledge of rational self-interest– anticipate betrayal and, therefore, does not participate in the transaction as $-\beta < 0$. ■

Remark 3 *As a matter of fact, as betrayal is a dominant strategy, the sequential character of the game does not affect the outcome of the Basic Trust game when played amongst RATs.*

At the heart of this negative outcome lies a credibility problem. Even though the self-interested trustee would be happy to forego the opportunistic incentives he faces in exchange for the truster's trust, his incapacity to commit his honourment to such placement, prevents the implementation of the Pareto Superior

outcome (place trust, honour trust if placed). Therein lies the credibility problem which is at the heart of the social inefficiency that characterises the result presented above.

However, it has to be taken into account that this prediction is built on the assumption that agents will maximise their material payoffs irrespective of the implicit obligations that might emerge in certain contexts. The advice of transaction-cost economics to overcome the non-cooperative outcome that emerges from opportunism[5] is to design a structure of credible commitments that protect transactions threatened by its hazards. However, it has to be borne in mind that, once the commitment problem is fully solved, a Relation of Trust ceases being one of trust.

In a number of possible transactions neither third party enforcement nor the long-term structure of incentives is there to motivate opportunistic agents to honour trust if it is placed in them. Does this mean that these transactions would not take place as eventual trusters would predict betrayal? The answer will depend on whether or not we stick to the assumption that trustees behave in an opportunistic fashion. If this was the case, i.e., if the whole population of trustees' were composed of opportunistic agents, one-shot transactions or repeated but unrecorded transactions (recurrent transactions) which possess the monetary structure incentives of the BT game, would not take place. However, if a behavioral disposition towards honouring were expected in relations of trust, the prediction of the negative would cease to be valid. In the next section, evidence which support the fact that an important proportion of individual agents honour trust when trust is placed in them despite the existence of monetary incentives for betraying it, will be presented. In the following section, it will be argued that this positive response can be understood within the framework provided by formal theories of reciprocity.

3 Social Pressure over Trustworthiness

This section explores the link between trust emergence between members of different communities and the emergence of a social norm in the group whose trustworthiness is required. Our aim is therefore to suggest a route to endogenise group reputation in a Relation of Trust where agents interact as anonymous members of their respective communities.

As stated before, either in a static environment or in a dynamic where agents from different groups cannot engage in repeated interaction to create an isolated transaction, investment reputation mechanisms are puerile. However, if mutual

[5] Williamson [20] distinguishes between "self-interest-seeking" and "opportunistic behaviour". Opportunism, in his approach, appears as a special case of self-interest-seeking. In contrast with simple self interest, according to which economic agents will continuously consult their own preferences but will reliably discharge all covenants, opportunistic agents are given to self-interest-seeking with guile. In our terminology, opportunism should be equated to material self-interest maximisation. However, for us a non-opportunistic agent will also fulfil the implicit obligations that emerge in certain structures of interaction.

enforcement is available, even opportunistic agents, as it happens in the individual reputation case, could behave in a trustworthy way in order to avoid the costs of the sanctions inflicted by his peers. Understanding group reputation as a public good, what we have is that peer-to-peer enforcement prevents opportunistic agents to free-ride on moral ones.

In the model presented in this section, the interaction between members of different communities takes the form of a recurrent asymmetric game. In such a set up both a) the evolution of a social norm which enforces trust honouring within the trustees' community and b) between-communities trust emergence will be analysed as inter-dependent processes. In section (3.3), in a strategic framework, it will be shown how group reputation could allow trust to emerge if exogenously motivated by moral dispositions. In section (3.4), the common knowledge assumption of the moral structure of the trustees' community and the unbounded rationality assumption will be relaxed to explore the implications of the existence of a payoff-dependent imitation process within each community.

3.1 The Group Reputation game (GR game)

Interaction and Sequential Structure of the GR game There are two disjoint populations (or communities): the one composed of the trusters and the one composed of the trustees. In every period dt randomly paired pairs –constituted by one truster and one trustee– recurrently play the Basic Trust game described in section (2). After all the BT games have been played, in the same period dt the "enforcement game" which is a simultaneous game played among all the members of the trustees' community, takes place. The game played in every period dt is then constituted by the games played between the randomly constituted pairs, who will play the BT games, and by the enforcement sub-game played among the trustees. The game which involves both the trust games and the enforcement sub-game will be referred to as the Group Reputation stage game and the game played recurrently simply as the Group Reputation game (GR game).

The Enforcement sub-game Every trustee, in the *enforcement game*, decides whether or not to monitor the actions of a proportion $\eta \in [0, 1]$ of the whole population of trustees. The monitoring process reveals the action of his colleagues in the trust game. If he chooses to monitor a group of colleagues, he has to decide whether or not to punish each one of them. Information about the monitored trustee's behaviour in previous rounds will not be available for the monitoring agent. The lack of a historical record might be associated with a small probability for the same trustee to be monitored in different rounds by one of his colleagues and/or bounded memory.

Enforcement costs for a trustee i, irrespective of his behaviour towards the trusters, are $\tau k_i + c$.; c is the fixed cost associated with the monitoring process and τk_i is the marginal cost of punishing a proportion k_i of the colleagues monitored by him.

The disutility associated with the cost inflicted by the enforcers is equal to $-\widehat{\lambda} l_i$, where l_i is the proportion of colleagues who punish the trustee i (that is, if the whole community punishes the trustee i, he suffers a disutility of $-\widehat{\lambda}$). It is convenient to define $\lambda = \widehat{\lambda}\eta$ which is the expected punishment damage for an agent who would be punished by every trustee who monitors him if each trustee in the community decided to perform monitoring.

Regarding parameter restrictions in the game as a whole, we assume $\delta - \lambda < \gamma$, that is, it is not optimal to betray trust in the worst case scenario, i.e., when the rest of your colleagues are sanctioning betrayal. In addition, we assume $\delta - \beta < \gamma - c + \alpha$ which implies that provided compensation mechanisms are allowed, the outcome (trust is placed, trust is honoured) is Pareto Superior to (trust is placed, trust is betrayed) even when fixed costs associated with monitoring are taken from the trustee's payoff if the cooperative outcome is achieved.

3.2 The GR game played amongst RATs

In this section, we assume that individuals of both communities maximise their monetary payoffs and this fact is *Common Knowledge*. Under these assumptions, we have the following outcome:

Proposition 4 *If the GR game is played amongst rational individuals driven by material self-interest (RATs) and this fact, the structure of the game and all the parameter values are common knowledge, then the unique sub-game perfect Nash equilibrium (SPNE) of the GR game is: every trustee betrays trust if trusted and does not monitor any colleague; every truster distrusts. That is, the outcome of the game is distrust.*

Proof. In the trustees' community, there are four classes of strategies:

$$\begin{array}{rcl} FM & : & \text{honour trust if trusted and monitor his colleagues.} \\ F & : & \text{honour trust if trusted and do not monitor his colleagues.} \\ BM & : & \text{betray trust if trusted and monitor his colleagues} \\ B & : & \text{betray trust if trusted and do not monitor his colleagues} \end{array}$$

Expected payoffs associated with the previous strategies for a trustee i are given by

$$\begin{array}{rcl} \pi_i^{FM} & = & \gamma - \lambda l_f - \tau k_i - c \\ \pi_i^{F} & = & \gamma - \lambda l_f \\ \pi_i^{BM} & = & \delta - \lambda l_b - \tau k_i - c \\ \pi^{B} & = & \delta - \lambda l_b \end{array} \quad (1)$$

where l_f and l_b are the proportion of monitoring trustees that punish honourers and betrayers respectively. As $\pi_i^{FM} < \pi_i^{F}$ and $\pi_i^{BM} < \pi^{B}$, neither monitoring

nor punishment will take place, i.e., $q_{FM} = q_{BM} = 0$. This, in turn, implies that $l_b = l_f = 0$ which makes it optimal to betray trust if trusted and not to monitor. In the trusters community, expected payoffs with the two strategies available, trust (T) and distrust (D), are given by

$$\begin{aligned} \pi^T &= \alpha(q_{FM} + q_F) - \beta(1 - q_F - q_{FM}) \\ \pi^D &= 0, \end{aligned}$$

As the trusters anticipate that $q_{FM} = q_F = 0$, the expected utility of trusting is $\pi^T = -\beta$; the truster's optimal strategy is to distrust. Common knowledge of rationality and self-interest drives group reputation, measured as the trustee's estimated probability of trust honouring, to the lowest possible level: zero. The Common Knowledge assumption implies that group reputation coincides with the actual behaviour of the trustees' community, that is, with the effective level of trustworthiness. ■

Thus, given a population of trustees homogeneously compounded by self-interest driven agents, the prediction of standard game theory is distrust between groups –the argument presented in Kandori [15], where an extension of the folk theorem is presented to justify the emergence of a social norm on the basis of self-interest, is not valid in our setting. His result is based on the record of the history of play, which is assumed to be unavailable in our model. Notice that social behaviour in the GR game in this scenario is exactly the same than in the BT game where mutual punishment devices within the trustees' community are not available (see Proposition 2).

3.3 The Multiple-type Group Reputation game (MGR game)

Consequentialist models of fair behaviour fall in two different categories, those based on distributional preferences and those which take into account the intentions of agents in a particular context. Fairness models based on *inequity aversion* such as those developed by Fehr and Schmidt [8] and Bolton and Ockenfels [3], fall in the former group. Alternatively, models as those developed by Rabin [17] and Dufwenberg and Kirchsteiger [6] explain moral behavior as motivated by *reciprocity*, fall in the latter category. Reciprocators, rather than having an optimal distribution on mind guide their actions to correspond the nature of the actions oriented towards them.

Under our account of relations of trust, the nature of the moral norm that obligates the trustee to honour trust can be understood as a particular case of reciprocity. It is the voluntary exposure –understood as a nice action– that produces in a reciprocator the disposition to honour trust. Notice that in the absence of an exit option for the truster, the reciprocity argument does not apply as the trustee has no form to qualify the intentions of the truster. Experimental evidence supporting the existence of this disposition can be found in Snijders and Keren [19] who investigated a trust game with a parameter structure which fulfils the conditions of the BT game studied here.

If the decision of the truster of whether or not to honour trust is affected by positive reciprocity effects; on similar ground it could be argued that the decision of whether or not to spend resources sanctioning betrayers is affected by negative reciprocity effects. Supporting this claim, Fehr and Gachter [7] present evidence of reciprocity effects in a public good provision game. The relevant result for our setting is the one obtained from the "Stranger" treatment with punishment opportunities. In a two-round public good game (in the second round punishment was available), costly punishment was implemented by the experimentees (in an inverse proportion to the amount of the public good provided by the punishee). Furthermore, the availability of punishment increased the total amount of the public good provided by the group with respect to the "Stranger" treatment without punishment opportunities.

In this scheme, for a high enough sensitivity to reciprocity effects agents would honour trust in the trust game and punish betrayers in the enforcement game. In the former case, the action of trust understood as a nice action could motivate honouring of trust and; in the latter case, the action of betrayal understood as a nasty action, the lack of contribution to a public good: group reputation, could trigger punishment. In this section, we consider a stylised model of the moral structure of the trustees' community where $\widehat{q}$ denotes the proportion of reciprocally motivated agents who would honour trust in the first part of the GR game and sanction his colleagues in the Enforcement sub-game. The rest of the population, i.e., the proportion $(1-\widehat{q})$ is assumed to be self-interest driven –thereby the stylised character of the model: we assume moral dispositions to be dichotomous, either the agent is motivated by reciprocity effects or not. In addition, notice that we implicitly assume that agents who honour trust, but do not sanction their betrayer colleagues are not observed. Thus, in this model, Nature decides the two possible types that characterise the population of trustees: "Reciprocators" and "RATs" with probabilities $\widehat{q}$ and $(1-\widehat{q})$ respectively. As usual, in any interaction between a truster and a trustee, the trustee's type is not observable to the truster.

In this setting, the outcome can be characterised by the vector (q,p) where p denotes the proportion of trusting agents; and q, the proportion of reciprocally motivated trustees plus the proportion of trustees whose honouring is induced by the deterrence effect of the enforcers. I.e., $\widehat{q} \leq q$.

In the following proposition, the predicted outcome of the game is presented for different values of $\widehat{q}$.

Proposition 5 *In any PBE of the MGR game, reciprocators honour trust and sanction betrayers. If*

$$\widehat{q} \geq \frac{\delta-\gamma}{\lambda}, \tag{2}$$

the PBE of the MGR game is $(q,p)=(1,1)$*, that is, complete and perfect trust will emerge; if*

$$\frac{\beta}{\alpha+\beta} \leq \widehat{q} < \frac{\delta-\gamma}{\lambda}, \tag{3}$$

the PBE is $(q,p)=(\widehat{q},1)$*, that is complete but imperfect trust will emerge; and*

otherwise, i.e., if

$$\widehat{q} < \min\{\frac{\delta-\gamma}{\lambda}, \frac{\beta}{\alpha+\beta}\}, \tag{4}$$

the PBE is $(q,p) = (\widehat{q},0)$,*that is, the outcome of the game is distrust.*

Proof. Results obtained in the proposition above are valid under the assumption that the parameters, including the rate of enforcers, are common knowledge. In such a set up, group reputation, exogenously given by $\widehat{q}$, determines whether the efficient outcome can be achieved. Due to the Common Knowledge assumption, group reputation coincides with the effective trustworthiness (and enforcing) level of the trustees' community.

If the condition expressed in equation (2) holds, complete and perfect trust will emerge as the proportion of enforcers makes it unprofitable to betray trust. Expected payoffs for the self-interested trustees associated with the actions of honouring and betrayal are given by

$$\begin{aligned} \pi^F &= \gamma \\ \pi^B &= \delta - \lambda\widehat{q}, \end{aligned}$$

and betrayal will be deterred if and only if $\pi^E \geq \pi^B$, which can be rearranged as condition (2).

Conversely, if the condition expressed in equation (3) holds, it is still in the interest of the opportunistic trustee to betray trust. However, the optimal strategy on the truster's side is to trust. It is noteworthy that under this scenario it is not the deterrant power of social pressure that triggers the emergence of trust. In this case, it is the group reputation of the trustees' community – given the potential gains/losses faced by the trustees– what makes profitable in expected terms to participate in the transaction. As in $(1-\widehat{q})$ proportion of the transactions this trust will be betrayed by self-interested individuals we refer to the outcome as complete but imperfect trust. Finally, if the proportion of reciprocally motivated agents is not high enough either to deter betrayal or to make it profitable to place trust, the outcome of the process is distrust. ■

From this analysis, it can be understood how once a sanctioning norm is internalised, its deterrence effect might sustain a moral norm of behaviour even when this norm has not been internalised by the entire community. If this is the case, however, we have to bear in mind that a moral norm must be non-outcome oriented to qualify as such. Thus, when condition (2) holds and it is in the interest of a self-interested trustee to honour trust, we should refer to this pattern of behaviour as *community enforcement*. In this case, community enforcement replaces the role of perfect third party enforcement. Notice that in the behavioural model, in opposition to the case where only self-interest is assumed to drive the whole population of the trustees, the availability of sanctioning does change the prediction of the outcome. Without punishment, complete trust would emerge only for $\widehat{q} = 1$, condition which is lessened to $\widehat{q} \geq \frac{\delta-\gamma}{\lambda}$ in the GR game. This is what allows the community itself to operate as an outcome-improvement device in problems of trust.

3.4 Social Learning in the GR game

In this section a social learning model based on the imitation of successful agents in material terms which develops in parallel in the two communities involved in the GR games, is considered. In this setting, therefore we explore the possibility that a result like the one obtained in the previous section could be understood as a by-product of a social learning process.

In our social learning setting, we assume that the initial behaviour in the trustees' community is taken as the initial conditions of a dynamic process. Whereas initial behaviour in the trustees' community is an expression of the initial "moral structure of the community"; behaviour in the trusters' community should be understood as a measure of initial perceived group reputation of the community they will interact with. Thereby, we configure a particular setting in which trustees do not have intrinsic moral dispositions, as they are ready to switch to more successful strategies, and trusters do not have a definite perception of the moral qualities of the community of trustees. In this scheme, we want to explore whether the somewhat ad hoc result obtained in the previous section could be understood as the by-product of a social learning process. Specifically, we assume that:

1. Agents are initially programmed to play a particular strategy. Trusters are initially programmed either to trust the trustee they face in an eventual transaction or distrust him. The agents who play the former strategy will be referred to as (T)rusting trusters and the ones who play the latter strategy as (D)istrusting trusters. The trustees, on the other side, will be assumed to be pre-programmed either to (i) betray trust if trusted and do not punish any betrayer or (ii) honour trust if trusted and punish betrayers who are trusted. The agents who play the former strategy will be referred to as (B)etrayers and the ones who play the latter strategy, the strong reciprocators, as (E)nforcers. The vector (q, p), as in the previous section, characterise the state of the two populations. It is assumed that (q_0, p_0) characterises the initial dispositions of the two communities.

2. In every period of time, agents of each population randomly learn, with some noise, the payoff obtained by one of their peers. They use this information to review their own strategy against the observed one. If they perceive that this other agent is performing better than they are, they will change to the strategy played by him; otherwise, they will stick to their original strategy.

3. Review rates are equal to one in each population and both populations have the same distribution of noise.

We denote a state of the subpopulation of the trusters by p (the proportion of trusting agents) and a state of the subpopulation of trustees by q (the proportion of enforcers). Because of the relaxation of the rationality assumption, the concepts of effective trustworthiness level and group reputation should be reconsidered. In the strategic framework, due to the common knowledge assumption,

both levels coincided and were associated with q. In this new framework q can also be understood as a measure of the effective trustworthiness level of the trustees' community. However, q will not necessarily coincide with group reputation anymore. Rather, group reputation, should be linked to p. A higher level of trusting agents would be the expression of a high group reputation of the trustees community as perceived in the trusters' community.

From the assumptions presented above, the imitation dynamics of this two-subpopulation model –following Weibull and Bjornerstedt [2]– can be represented by the following system of differential equations

$$\frac{dp}{dt} = \left(\pi_p^T - \overline{\pi}^p\right) p \tag{5}$$

$$\frac{dq}{dt} = (\pi_q^E - \overline{\pi}^q)q, \tag{6}$$

which constitute a mere re-scaling of the two-subpopulations replicator dynamics.

Profits associated with each strategy are given by

$$\pi_p^T = \alpha q - \beta(1-q) \tag{7}$$

$$\pi_p^D = 0 \tag{8}$$

in the trusters' population, and by

$$\pi_q^E = \gamma p - \tau(1-q) - c \tag{9}$$

$$\pi_q^B = \delta p - \lambda q \tag{10}$$

in the trustees' population; and population average payoffs are given by

$$\overline{\pi}^p = (\alpha q - \beta(1-q))p + 0 \times (1-p) \tag{11}$$

$$\overline{\pi}^q = (\delta p - \lambda q)(1-q) + (\gamma p - \tau(1-q) - c)q \tag{12}$$

substituting (7) and (11) in (5),and substituting (9) and (12) in (6), we end up with the following dynamic system

$$\frac{dp}{dt} = [\alpha q - \beta(1-q)](1-p)p \tag{13}$$

$$\frac{dq}{dt} = [q(\lambda+\tau) - p(\delta-\gamma) - \tau - c]\,(1-q)\,q. \tag{14}$$

The only candidates for asymptotic stability are the pure strategy Nash equilibria (when players are assumed to be driven by self-interest), since mixed strategy NE of asymmetric evolutionary games are not asymptotically stable under the replicator dynamics (see Samuelson and Zhang [18]). Pure strategy Nash equilibria are $(p_1, q_1) = (1, 1)$, provided $\delta - \lambda < \gamma - c$ and $(p_2, q_2) = (0, 0)$.

Proposition 6 *The fixed point* $(p_1, q_1) = (1, 1)$ *is asymptotically stable if and*

only if

$$\delta - \lambda < \gamma - c \tag{15}$$

and $(p_2, q_2) = (0, 0)$ *is asymptotically stable for all parameter values.*

Proof. In order to analyse stability of any fixed point (p_i, q_i), with $i \in \{1, 2\}$, we linearise the system around each point. Provided that a fixed point is *hyperbolic* (i.e., every eigenvalue of the Jacobian of the linearised system evaluated at (p_i, q_i) has a non-zero real part), by the Hartman-Grobman Theorem, it will be topologically equivalent to the fixed point at the origin of the linearisation of the dynamic system

Stability Analysis of $(p_1, q_1) = (1, 1)$:

The Jacobian evaluated at $(1, 1)$ is given by

$$J_1 = \begin{pmatrix} -\alpha & 0 \\ 0 & -\gamma + c + \delta - \lambda \end{pmatrix}. \tag{16}$$

Asymptotic stability requires that the Jacobian above is negative definite which requires in turn that:

$$\delta - \gamma + c - \lambda < 0 \tag{17}$$

Thus, if and only if the condition expressed in equation (17) holds, $(p_1, q_1) = (1, 1)$ is asymptotically stable.

Stability Analysis of $(p_2, q_2) = (0, 0)$:

The Jacobian evaluated at $(0, 0)$ is given by

$$J_2 = \begin{pmatrix} -\beta & 0 \\ 0 & -(\tau + c) \end{pmatrix}$$

As the parameters of the model are defined as positive, negative definiteness is guaranteed, and the fixed point $(p_2, q_2) = (0, 0)$ is always asymptotically stable.

That, is depending on the initial conditions of the system, the outcome of the process will be either distrust or complete and perfect emergence of trust. ■

The intuition behind the stability of (p_1, q_1) is the following. For any slight deviation from this point in the population of trusters, agents who are not placing trust will be worse-off than the ones who are doing so. In the population of trustees, betrayers will do badly as the high amount of enforcers make this strategy unprofitable (provided that the minimum difference in expected payoff between an enforcer and a betrayer is positive). In the case of (p_2, q_2), for any slight deviation from this point in the population of trusters, agents who are placing trust will be worse-off than the ones who are not doing so because they would face the costs associated with betrayal; and, in the population of trustees, enforcers will do worse as they would have to face the fixed cost c associated with the enforcement process. It is important to point out that in a set-up where

punishment was not available, the GR game would have only one asymptotically stable equilibrium: distrust.

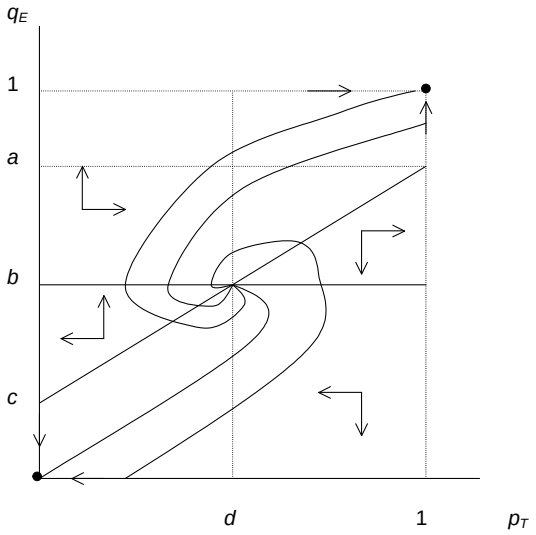
From the stability analysis, we obtained that there are two evolutionary equilibria: distrust and complete trust. That is, depending on the initial conditions of the system (p_0, q_0), trust might or might not emerge. The initial level of enforcers q_0 measures the initial effective trustworthiness level of the subpopulation of trustees; and the initial level of trusting agents p_0, the initial group reputation of the trustees' community as perceived by the members of the trusters' community. What follows is the relationship between the initial conditions for the emergence of trust and the parameter values which determine the structure of incentives of the GR game.

The isoclines of the system are given by

$$\frac{dp}{dt} = 0 \rightarrow q = \frac{\beta}{\alpha+\beta} \tag{18}$$

$$\frac{dq}{dt} = 0 \rightarrow q = \frac{c+\tau}{\lambda+\tau} + \frac{\delta-\gamma}{\lambda+\tau}p \tag{19}$$

The phase plane diagram is plotted below



where $a = \frac{\delta-\gamma+\tau+c}{\tau+\lambda}$, $b = \frac{\beta}{\alpha+\beta}$, $c = \frac{(c+\tau)}{(\tau+\lambda)}$ and $d = \frac{\beta(c-\lambda)+\alpha(\tau+c)}{(\alpha+\beta)(\delta-\gamma)}$. The graph shows the case characterised by $\frac{c+\tau+\delta-\gamma}{\tau+\lambda} \leq \frac{\beta}{\alpha+\beta} \leq \frac{(c+\tau)}{(\tau+\lambda)}$.

Based on the qualitative analysis of the trajectories depicted above, we establish a sufficient condition for the complete and perfect emergence of trust.

Proposition 7 *Given $q_0 > 0$ and $p_0 > 0$, a sufficient condition for the complete and perfect emergence of trust is*

$$q_0 > \frac{\delta - \gamma + \tau + c}{\tau + \lambda} \tag{20}$$

Proof. The condition for the stability of the complete trust outcome: $\lambda > \delta - \gamma + c$ guarantees that the minimum size of the basin of attraction of the fixed point associated with complete and perfect emergence of trust $(p_1, q_1) = (1, 1)$ is non empty since it implies that $q(p = 0) = \frac{(c+\tau)}{(\tau+\lambda)} \in (0, 1)$ and $q(p = 1) = \frac{\delta-\gamma+\tau+c}{\tau+\lambda} \in (0, 1)$. When the condition expressed in equation (20) holds, we have $\frac{dp}{dt} > 0$ and $\frac{dq}{dt} > 0$,.i.e., the proportion of both trusting trusters and enforcing trusters is growing. If the boundary associated with $p = 1$ is reached, we have $\frac{dq}{dt} > 0$ if and only if $q > \frac{\delta-\gamma+\tau+c}{\lambda+\tau}$ (replace $p = 1$ in equation 14) which is guaranteed as the boundary will only be reached for that interval; and, if the boundary associated with $q = 1$ is reached, we have $\frac{dp}{dt} = \alpha(1-p)p > 0$ (replace $q = 1.$ in equation 13). ■

One of the implications of Proposition (7) relates to the fact that trust emergence does not require a minimum level of initial group reputation. Provided that the effective level of trustworthiness is high enough, all that is required in terms of group reputation is that a positive proportion of trusters do place trust in the initial round of the recurrent game. This striking result tells us about the social relevance of the existence of a group of mavericks willing to expose themselves to other agents in a novel interaction structure. That is, in this framework, not only trustworthiness becomes a pubic good, individual trusting trusting dispositions also acquire a social dimension.

4 Discussion

In section (3.2), under the assumptions of rationality, self-interest and common knowledge thereof, it is shown that the prediction of the GR game is "distrust". As is shown in section (3.3), the previous result relies heavily on the self-interest assumption. Assuming a hybrid community of trustees where a proportion of the members of this community is reciprocally motivated, it was shown how complete trust could emerge as the result of the deterrence effect associated with a high proportion of those kind of agents. For a low proportion of reciprocally motivated agents, the predicted outcome was the same as the one obtained from the standard approach: distrust. In an intermediate interval for the sensitivity to reciprocal effects, imperfect emergence of trust was predicted. In section (3.4), using a payoff-dependent imitation approach, similar results were obtained to the ones obtained from the behavioural analysis (where reciprocity effects where taken into account). To link the results obtained from the imitation dynamics and the behavioural model based on reciprocity effects, it is illuminating to view the initial conditions of the dynamic system analysed in section (3.4) as initial moral dispositions of the trustees' community and its perception in the trusters

community.

The analysis developed in this paper shows how community governance goes beyond the domain of internal problems of collective action via the regulation of the relation of the community with the rest of the society. Consider the problem of a national health service where doctors see patients in a rotative fashion, i.e., the probability that one doctor will see the same patient in a finite period of time is close to zero. As the exercise of the profession requires specific knowledge, patients will have inferior information compared to the providers regarding the quality of the service provided. Therefore, opportunistic incentives arise on the doctors' side, let us say, to offer a service of lower quality at a lower effort (see Iversen and Luras [12] for an empirical analysis of the role of professional norms in medical practice). Patients cannot identify medical malpractice on the spot, but can infer it after sharing their experiences with other patients. Because of the recurrent character of the interaction, relations become anonymous, encapsulated long-term relations cannot be created and, therefore, personal enforcement cannot work. However, if a professional norm emerges in the community of doctors, complete and perfect trust could emerge. Such a professional norm should be of the form: offer a proper medical service and punish those colleagues who do not offer a proper service. As we saw in section (3.4), trust might emerge even when the initial reputation of the doctors' community is low. Monitoring of colleagues' behaviour is possible because of the inter-dependent character of medical specialities (when a psychiatrist refers one patient to a neurologist, the latter, at a certain time cost, can monitor the former's service). Punishment could take the form of social ostracism, e.g., stop inviting him to conferences or other activities associated with the profession. If a private system co-exists with the national public system, and part of the clients in the private system are sourced from the public system, the punishment could become more effective: a neurologist could stop referring patients to the private clinic of a particular psychiatrist if he observes the inadequate professional behaviour of the psychiatrist in the National Service. Another possible application of the model is the between-groups trust problem which arises from inter-dependent academic activity (e.g., anthropologists using field results from ethnologists).

References

[1] Annette Baier. Trust and antitrust. *Ethics*, 96(2):231–60, 1986.

[2] Jonas Bjornerstedt and Jorgen Weibull. Nash equilibrium and evolution by imitation. In Enrico Colombatto Kenneth Arrow and Christian Schmidt, editors, *The Rational Foundations of Economic Behavior*, pages 155–71. Macmillan, London, 1996.

[3] Gary Bolton and Axel Ockenfels. Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90, 2000.

[4] James Coleman. *Foundations of Social Theory*. Harvard University Press, Cambridge, Massachusetts, 1990.

[5] Partha Dasgupta. Trust as a commodity. In Diego Gambetta, editor, *Trust: making and breaking cooperative relations*, pages 49–72. Blackwell, New York, 1988.

[6] Martin Dufwenberg and Georg Kirchsteiger. A theory of sequential reciprocity. *Working Paper Tilburg University*, 1998.

[7] Ernst Fehr and Simon Gachter. Cooperation and punishment in public good experiments. *American Economic Review*, 90, 2000.

[8] Ernst Fehr and Klaus Schmidt. A theory of fairness, competition, and co-operation. *Quarterly Journal of Economics*, 118, 1999.

[9] Alvin Gouldner. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25, 1960.

[10] Russell Hardin. Trust. In Peter Newman, editor, *The New Palgrave Dictionary of Economics and the Law*. Macmillan, Basingstoke, 1998.

[11] Russell Hardin. Conceptions and explanations of trust. In Karen Cook, editor, *Trust in Society*, pages 3–39. Russell Sage, New York, 2001.

[12] Tor Iversen and Hilde Luras. Economic motives and professional norms: The case of general medical practice. *Journal of Economic Behavior and Organization*, 43, 2000.

[13] Harvey James. The trust paradox: A survey of economic inquiries into the nature of trust and trustworthiness. *Journal of Economic Behavior and Organization*, 47(3):291–307, 2002.

[14] John Dickhaut Joyce Berg and John McCabe. Trust, reciprocity and social history. *Games and Economic Behavior*, 10(1):122–42, 1995.

[15] Michimiro Kandori. Social norms and community enforcement. *Review of Economic Studies*, 1992.

[16] David Kreps. Corporate culture and economic theory. In Kenneth Shepsle and James Alt, editors, *Perspectives on Positive Political Economy*, pages 90–143. Cambridge University Press, New York, 1990.

[17] Matthew Rabin. Icorporating fairness into game theory and economics. *American Economic Review*, 1993.

[18] Larry Samuelson and J. Zhang. Evolutionary stability in assymmetric games. *Journal of Economic Theory*, 57, 1992.

[19] Chris Snijders and Gideon Keren. Determinants of trust. In Ido Erev David Budescu and Rami Zwick, editors, *Games and Human Behaviour*, pages 355–385. Lawrence Erlbaum, Mahwah, NJ, 1999.

[20] Oliver Williamson. *The Economic Institutions of Capitalism : Firms, Markets, Relational Contracting*. Free Press, New York, 1985.

[21] Oliver Williamson. Calculativeness, trust and economic organisation. *Journal of Law and Economics*, 36, 1993.