An N-player Trust game: comparison of replicator and logit dynamics

I. N. Towers [a], **B. O'Neill** [a], **H. S. Sidhu** [a], **Z. Jovanoski** [a], **K. Merrick** [b], **and M. Barlow** [b]

[a]*School of Physical, Environmental and Mathematical Science, UNSW Canberra*
[b]*School of Engineering and Information Technology, UNSW Canberra*
Email: i.towers@adfa.edu.au

Abstract: Trust and trustworthiness are fundamental aspects of society. Trust acts as the glue which holds a social system together. The greater the levels of trust, the greater the stability of the society. Further, trust acts as a simplifying heuristic or complexity reduction mechanism. Once trust is established between members of a society it makes possible certain types of exchanges which were not before or greatly simplifies mechanics of existing transactions.

The phenomenon of trust is often studied in a game-theoretic environment known as the trust game or investment game. In this work we consider a variant of the trust game played by N players where N is sufficiently large such that the population of players can be modelled as a continuous quantity. Players may choose the role of trustor, honest trustee, or dishonest trustee continuously throughout the game and earn the associated pay-offs. Decisions are made via two different strategy revision functions: replicator and logit dynamics.

We show that when players base their strategy decisions upon mimicking the successful strategies of players they encounter (replicator dynamics) there is no stable equilibrium in the model with non-zero population proportion of trustors. In contrast, when players choose the strategy with the greatest pay-off in their current situation, but in an environment without perfect knowledge of what is best, a stable equilibrium with trusting players exists.

This work has application in the area of trusted autonomy. Trusted Autonomy refers to two or more interacting and self-governed autonomous intelligent systems (including humans) where one side of the interaction is willing to delegate a task that will make it vulnerable to other parties in the interaction who are willing to accept and can autonomously perform the task. As such these results could inform design decisions of these autonomous systems.

Keywords: *Game theory, trusted autonomy, trust game, investment game, logit dynamics*

1 THE TRUST GAME

Trust and trustworthiness are fundamental aspects of society. Trust acts as the "glue" which holds a social system together. The greater the levels of trust, the greater the stability of the society. Further, trust acts as a simplifying heuristic or complexity reduction mechanism. Once trust is established between members of a society it makes possible certain types of exchanges or greatly simplifies mechanics of existing transactions.

As critical as trust is to society, its study is most commonly found in the fields of economics, psychology and neuroscience. It is less common in theoretical biology and ecology while the concept of trust is absent in the literature of evolutionary computation (Abbass et al., 2016).

When trust is studied it is often in terms of a game-theoretic framework especially amongst economists known as the "trust game" or "investment game". The now classical version of the trust game was presented in Berg et al. (1995). The game consists of two players: A (trustor) and B (trustee) who are both given the same amount of money. Player B automatically keeps their initial funds while player A must decide how much (if any) of their funds will be transferred to B. A knows that any funds transferred will be increased by a factor of, say, 3 and then B will decide how much (if any) of the increased pot will be transferred back.

Economic theory allows for any preference ordering over outcomes (or gambles of outcomes), but under some basic rationality requirements these can be manifested in utility functions that represent the preferences (see e.g., Jehle and Reny (2001, 5-18,92-112)). Simple models of behaviour for the two players in the trust game posit that each players preference ordering is entirely determined by his own funds and his utility is some increasing concave function of his own funds. In game theoretic applications players are assumed to be able to reason strategically about their opponents actions and so there is convergence to a Nash equilibrium, where each player acts with a strategy that is a best-response to his opponents strategy i.e., there is no unilateral incentive for a player to change his strategy.

This means that each player attempts to maximise the expected utility from his own funds, subject to weighting of risk, and each player plays a best response to his opponents strategy based on this optimising criterion. For a single play of this game, B acting rationally should keep all the money he is entrusted with by A and give nothing back. This is his best-response to any strategy by A and so it is a dominant strategy. Since A is also rational he will realise that this is a dominant strategy for B and so he will choose not to invest any of his funds with B. Thus, in any anonymous single play of the game there is no rational reason to trust and all potential trust benefits of the transaction are lost (Manapat et al., 2013).

Fortunately the dismal predication of this simple economic analysis is not what Berg et al. (1995) (nor other experimenters that followed) saw in experiments. In practice, trustors choose to transfer funds with high probability and trustees choose to return substantial amounts (Berg et al., 1995; Engle-Warnick and Slonim, 2006; Neo et al., 2013). These results lead to questions of why do we trust and how does trust evolve in a society? The existing literature reports attempts to make inroads on these questions by adapting experimental aspects of the trust game and altering theoretical models of the situation. Games have been played as blocks of single-shots (Neo et al., 2013) or blocks of repeated games with the same two players (Engle-Warnick and Slonim, 2006). Concepts like reputation, punishment and moral codes have been incorporated experimentally (Manapat et al., 2013) and theoretically (Masuda and Nakamura, 2012; Rabanal and Friedman, 2015).

Tzieropoulos reviews the trust game from a neuroscience perspective and reports the importance of a number of factors in the outcome of experiments such as: past experience with the trust game; reputation formation and learning adaptations over repeated games; delays in learning the outcome of a set of games; restrictions on what amounts could be transferred in the game; and enforced delays in player decisions (Tzieropoulos, 2013).

The aim of this work is to extend the theoretical analysis of the trust game presented in Abbass et al. (2016). Therein the authors look at a pure trust game of N players who can change both roles and strategies via so-called *replicator* dynamics. In this work we will extend the existing model by considering the effects of imperfect knowledge on players strategy choices via the so-called *logit* dynamics (Rabanal and Friedman, 2015) and compare the resulting behaviour with that of the original model.

2 REPLICATOR DYNAMICS

In this section we consider the results of Abbass et al. (2016). Their version of the trust game is summarised in fig. 1 below. Every player in the game chooses a role: either trustor or trustee. The trustors are the source of "wealth" or "trust" within the game but they are reliant upon the trustees for something to come of that wealth. Players who choose to be trustees have one further choice to make: to be trustworthy or not. Each trustor

invests a unit of wealth with the trustees and the total investment is distributed equally amongst all the trustees. Trustworthy trustees achieve a return on investment of $2R_1$ (where $R_1 > 1$) and return half the wealth to the trustors. Untrustworthy trustees achieve a return on investment of R_2 (where $1 < R_1 < R_2 < 2R_1$) and keep all of it.

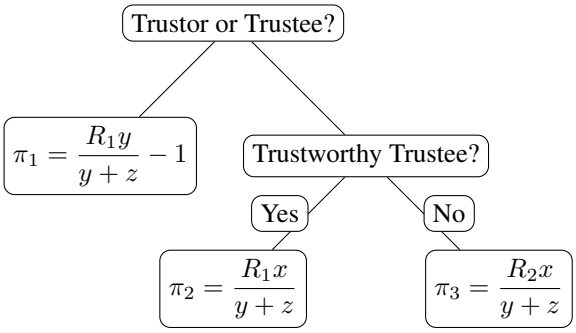

Figure 1. The pay-offs, π_j, for each of the strategies in the trust game.

When considering the N player single shot trust game the variables x, y, z are the numbers of trustors, trustworthy trustees and untrustworthy trustees respectively such that $x + y + z = N$. Implicit in the pay-offs of the game is that $y + z > 0$ i.e. that there is at least one trustee. In the case of no trustees then no game of trust is able to be played.

The game-theoretic analysis of the single shot trust game of Abbass et al. (2016) produces similar results as when applied to the game of Berg et al. (1995). For a single play of this game, the choice to be an untrustworthy trustee is the best-response to any strategy by the other players and so it is a dominant strategy. The Nash equilibrium of the game occurs when all players choose this strategy, and in this case there is no unilateral incentive for any player to change strategy. Although, playing as an untrustworthy trustee is rational for the individual it minimises the total wealth of the population. Abbass et al. show that wealth is maximised if $x = N - 1$ and $y = 1$ i.e. 1 player is a trustworthy trustee and the remainder are trustors. This wealth-maximising outcome is highly unstable because there is a large incentive for any player to change strategy to being an untrustworthy trustee.

It is worth remarking here on the implicit "production" process specified in this game. The game specifies that the production of investment returns is conducted by trustees, but the production of returns treats the process as being something that can be done by a single trustee without regard to scale. This means that only a single trustee is required to produce investment returns regardless of the scale of investment, and any further trustees are deadweight in this process. Moreover, the model also imposes a penalty in production for untrustworthy trustees so that it is implicitly assumed that the latter are somehow worse at investment than the trustworthy trustees. The reason for this is not specified. The wealth-maximising outcome of the game comes about because the setup of the game makes the existence of a trustee a necessary condition for a return on investment, but the return does not increase if the number of trustees increases (although it is affected by the proportion of trustees that are trustworthy).

In order to create an evolutionary model we assume that the total population becomes large i.e. $N \to \infty$. Following Sandholm (2009), the rate of change in the proportion of players choosing strategy i is

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} x_j \rho_{ji} \left(\pi(\mathbf{x}), \mathbf{x}\right) - x_i \sum_{j=1}^{n} \rho_{ij} \left(\pi(\mathbf{x}), \mathbf{x}\right) \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \cdots x_n)$, $\sum_{k=1}^{n} x_k = 1$, every element x_k is non-negative, and ρ_{ij} is proportional to the probability that a player utilising strategy i switches to strategy j. Equation (1) is a generic expression for the so-called mean dynamical behaviour of the strategy x_i.

We will assume a player only changes strategy if their opponent's strategy has a greater pay-off and they do so

with probability proportional to the function

$$\rho_{ij} = \begin{cases} x_j\left(\pi_j - \pi_i\right) & \text{if } \pi_j > \pi_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The function ρ_{ij} is known as the revision function. Substitution of eq. (2) into eq. (1) results in the replicator dynamic system of differential equations

$$\frac{dx_i}{dt} = x_i\left(\pi_i(\mathbf{x}) - \overline{\pi}(\mathbf{x})\right) \tag{3}$$

where $\overline{\pi}$ is the average pay-off of the population.

On substituting the pay-offs for the trust game into eq. (3), where now we define $\mathbf{x} = (x, y, z)$, and using the constraint $x + y + z = 1$ the evolutionary trust game with replicator dynamics is derived as,

$$\begin{aligned} \frac{dx}{dt} &= \frac{x^2}{1-x}\left[y\left(1-2R_1\right) + z\left(1-R_2\right)\right] + \frac{x}{1-x}\left[y\left(R_1-1\right) - z\right], \\ \frac{dy}{dt} &= \frac{xy}{1-x}\left[y\left(1-2R_1\right) + z\left(1-R_2\right) + R_1\right], \\ \frac{dz}{dt} &= \frac{xz}{1-x}\left[y\left(1-2R_1\right) + z\left(1-R_2\right) + R_2\right]. \end{aligned} \tag{4}$$

System (4) has two equilibrium points $\mathbf{x}^* = (x^\star, y^\star, z^\star)$ such that $\dot{x}\left(\mathbf{x}^*\right) = \dot{y}\left(\mathbf{x}^*\right) = \dot{z}\left(\mathbf{x}^*\right) = 0$. The first equilibrium is a stable line in the phase space $\mathbf{x}^* = (0, a, 1-a)$ where $a \in [0, 1]$. At this point all players are some form of trustee and the total wealth of the population is zero. The second equilibrium is, in general, unstable *unless* the system's initial condition is such that no player has chosen to be an untrustworthy trustee. In the case of this somewhat artificial starting point the equilibrium

$$(x^\star, y^\star, z^\star) = \left(\frac{R_1 - 1}{2R_1 - 1}, \frac{R_1}{2R_1 - 1}, 0\right),$$

is stable and the long term mix of the fraction of trustors and trustworthy trustees is dependent only on the parameter R_1. This is consistent with the concept behind the replicator dynamic — if there is no one already pursuing the untrustworthy trustee strategy then it's impossible for others to mimic it and thus the system (4) will not generate a non-zero z.

A more typical scenario for (4) is illustrated in fig. 2 where the initial condition is set as $(x(0), y(0), z(0)) = (0.9, 0.099, 0.001)$ and the parameters are set to $R_1 = 6$ and $R_2 = 8$. Even when a tiny fraction of the population initially utilise the untrustworthy strategy it quickly results in the vast majority of the population joining them and driving the trustor fraction towards zero as shown in fig. 2.

3 Logit Dynamics

The logit dynamics are derived using a different strategy revision function to the replicator dynamics. Assume the player utilising strategy i re-evaluates their strategy and will switch to the strategy j with probability ρ_{ij} given by

$$\rho_{ij} = \frac{e^{\pi_j/\eta}}{\sum_{k=1}^{n} e^{\pi_k/\eta}} \tag{5}$$

where the parameter $\eta > 0$ measures the difficulties in perceiving the best strategy and is often called the *noise level*. If η is large, choice probabilities under the logit rule are nearly uniform — the difficulty perceiving the best strategy is high so all strategies appear to be good choices. But if η is near zero, choices are optimal with probability close to one, at least when the difference between the best and second best pay-off is not too small. Further, as the numerator of eq. (5) does not depend on i this implies you are equally likely to transition to a new state regardless of your present state.

If we now substitute the logit revision function eq. (5) into eq. (1) then we produce the generic logit dynamic system of equations

$$\dot{x}_i = \rho_{ii} - x_i \tag{6}$$

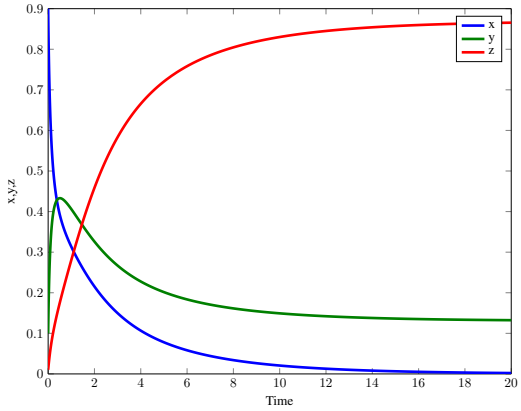

Figure 2. The dynamical behaviour of the population fractions when $R_1 = 6$, $R_2 = 8$ with non-zero $z(0)$.

We can make the generic system (6) specific to the trust game by substituting the particular pay-offs into eq. (5). The result is

$$\frac{dx}{dt} = \rho_{11} - x, \qquad \frac{dy}{dt} = \rho_{22} - y, \qquad \frac{dz}{dt} = \rho_{33} - z \tag{7}$$

where the ρ_i are defined as

$$\rho_{ii} = \frac{e^{\pi_i/\eta}}{e^{\pi_1/\eta} + e^{\pi_2/\eta} + e^{\pi_3/\eta}} \tag{8}$$

While appearing simpler than eq. (4) the logit system (7) is not generally tractable to find the equilibrium points and studying the stability via linear analysis. First we will consider two limiting cases when the system is tractable.

In the limit of large noise level $\eta \to \infty$ the players have no information to determine which is the best strategy to follow at a given instant. In this case all revision functions behave as $\rho_{ii} \to \frac{1}{3}$. Thus in the large noise limit eq. (7) decouples and is easily solved to give the exact solution

$$\mathbf{x} = \mathbf{x_o} e^{-t} + (1/3, 1/3, 1/3)^T , \qquad \mathbf{x_o} = \mathbf{x}\,(t = 0)\,. \tag{9}$$

Unsurprisingly, if players are unable to differentiate one strategy from another then rationally they will choose strategies with equal likelihood which must express itself in the long term as an equipartition of the population amongst the strategies.

In the opposite limit of perfect knowledge (no noise) $\eta \to 0^+$ then players always know the most beneficial strategy to pursue at any given moment and the revision functions tend to $\rho_{11} \to 0$, $\rho_{22} \to 0$ and $\rho_{33} \to 1$. Once again the system decouples in this limit and an exact solution can be derived

$$\mathbf{x} = \mathbf{x_o} e^{-t} + (0, 0, 1)^T . \tag{10}$$

Now the population rapidly reverts to the untrustworthy strategy from any initial condition. Ironically, lack of information produces an equilibrium of greater overall wealth than knowing perfectly which strategy has the greatest pay-off. Imperfect perception stops the players destroying the investment return in pursuit of personal gain.

In general for finite positive η, however, it's not possible to find a closed form solution for the equilibrium point. The existence of at least one mixed equilibrium point (i.e. $\mathbf{x} \neq \mathbf{0}$) can be made using index theory (Jordan and Smith, 2007, 89).

First, using the constraint $x + y + z = 1$, we can rewrite the pay-offs as

$$\pi_1 = \frac{R_1 y}{1 - x} - 1, \qquad \pi_2 = \frac{R_1 x}{1 - x}, \qquad \pi_3 = \frac{R_2}{R_1}\pi_2.$$

This allows us to eliminate z entirely from eq. (7). Next, consider the closed curve in the x, y plane bounded by the lines $x = 0$, $y = 0$, and $y = 1 - x$. This curve bounds the allowable values of x and y. Along the line $x = 0$ the system (7) reduces to

$$\begin{aligned} \frac{dx}{dt} &= \rho_1 - x = \frac{e^{\eta^{-1}(R_1 y - 1)}}{2 + e^{\eta^{-1}(R_1 y - 1)}} > 0, \\ \frac{dy}{dt} &= \rho_2 - y = \frac{1}{2 + e^{\eta^{-1}(R_1 y - 1)}} - y. \end{aligned} \tag{11}$$

The vector field (x', y') along $x = 0$ is always pointing rightwards. At the point $(0, 0)$ both x' and y' are positive while at $(0, 1)$

$$\frac{dy}{dt} = \frac{1}{2 + e^{\eta^{-1}(R_1 - 1)}} - 1 < 0$$

By similar argument, along $y = 0$ the vector field points upwards except at $x = 1$. In the limit $x \to 1$ we find $x' = -1$ and $y' = 0$. Looking at the vector field at the 3 vertices of the triangle it is clear that field has gone through a complete revolution and therefore has an index of 1 and therefore has at least one equilibrium point.

Numerically we can solve the system and calculate the eigenvalues at the equilibrium points. A single equilibrium point moves from the origin ($\eta \to 0$) and for small values of η the equilibrium is a stable spiral. Oscillations near the equilibrium are quickly lost with moderate increases in noise. Examples of times series and phase portraits follow in figs. 3 and 4.

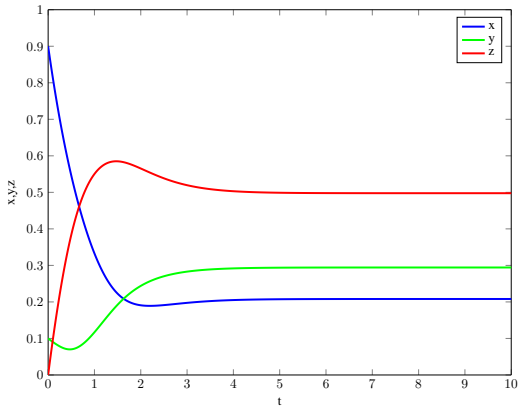

Figure 3. The dynamical behaviour of the population fractions when $R_1 = 6$, $R_2 = 8$ and $\eta = 1$.

4 Conclusion

Under a standard evolutionary model of the multiplayer trust game, if there are any untrustworthy trustees at the start of the game then there is convergence to the Nash equilibrium in which all players choose to be trustees and there is no investment. This is an unsurprising result — it occurs because the existence of untrustworthy trustees acts as a drain on the system which incentivises a shift to being a trustee. The same outcome is found in the variation of the trust game presented in Abbass et al. (2016).

Without any in-built "mis-perception" in the dynamics the players are able to see the state of the system and there is convergence to this degenerate outcome. However, when the players have limitations to their knowledge of the best strategy added by using an evolutionary form with logit dynamics and with a noise level parameter dampening the perceived profit then there is interference with the transition between states. This leads to a non-degenerate steady state involving a modicum of trust and positive, but sub-optimal, overall wealth.

The logit version of the Abbass et al. trust game can be seen as improvement over the original by the simple fact that trust exists in the system for finite noise level regardless of the initial population distribution amongst

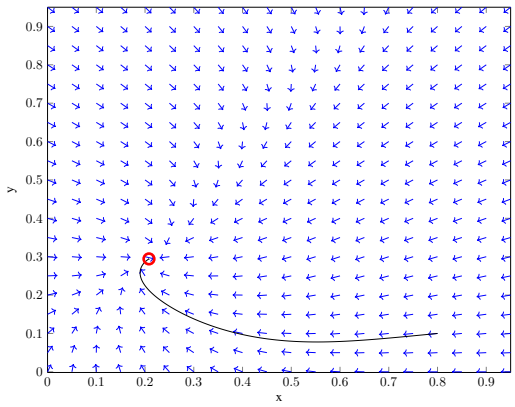

Figure 4. The phase space of trustors (x) versus trustworthy trustees (y) when $R_1 = 6$, $R_2 = 8$ and $\eta = 1$. The equilibrium point, represented by the red circle, is at $\mathbf{x}^\star = (x^\star, y^\star) = (0.2081, 0.2943)$. The eigenvalues of the Jacobian at $\mathbf{x}^\star$ are $\lambda_1 = -2.3685$ and $\lambda_2 = -1.5379$ indicating that $\mathbf{x}^\star$ is a stable node of the system.

strategies (i.e. it is not required that everyone be trustworthy from the outset). However, this trust comes at the expense of hobbling the players with imperfect knowledge of their best strategy at any given moment. The results of the model could be interpreted as trust arising from mistakes players make choosing strategies. More players are mistakenly choosing to be trustworthy. Alternatively, uncertainty of one's best play may produce trust between players as trust is more rational than making potentially costly mistakes due to imperfect knowledge.

REFERENCES

Abbass, H., G. Greenwood, and E. Petraki (2016). The n-player trust game and its replicator dynamics. *IEEE Transactions on Evolutionary Computation 20*(3), 470–474.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and Economic Behavior 10*(1), 122–142.

Engle-Warnick, J. and R. L. Slonim (2006). Learning to trust in indefinitely repeated games. *Games and Economic Behavior 54*(1), 95–114.

Jehle, G. and P. Reny (2001). *Advanced Microeconomic Theory* (2 ed.). Addison-Wesley series in economics. Addison-Wesley.

Jordan, D. W. and P. Smith (2007). *Nonlinear ordinary differential equations* (1 ed.). Oxford University Press.

Manapat, M. L., M. A. Nowak, and D. G. Rand (2013). Information, irrationality, and the evolution of trust. *Journal of Economic Behavior & Organization 90*, S57–S75.

Masuda, N. and M. Nakamura (2012). Coevolution of trustful buyers and cooperative sellers in the trust game. *PLoS ONE 7*(9), e44169.

Neo, W. S., M. Yu, R. A. Weber, and C. Gonzalez (2013). The effects of time delay in reciprocity games. *Journal of Economic Psychology 34*, 20–35.

Rabanal, J. and D. Friedman (2015). How moral codes evolve in a trust game. *Games 6*(2), 150–160.

Sandholm, W. H. (2009). Evolutionary game theory. In *Encyclopedia of Complexity and Systems Science*, pp. 3176–3205. Springer New York.

Tzieropoulos, H. (2013). The trust game in neuroscience: A short review. *Social Neuroscience 8*(5), 407–416.