Learning to trust, learning to be trustworthy

Berger, Ulrich

DOI:
10.57938/3138728f-1b09-40e0-a6be-ada83850c8df

Published: 01/01/2016

Document Version:
Publisher's PDF, also known as Version of record

Document License:
Unspecified

Department of Economics
Working Paper No. 212

Learning to trust, learning to be trustworthy

Ulrich Berger

January 2016

Learning to trust, learning to be trustworthy

Ulrich Berger*

December 11, 2015

Abstract

Interpersonal trust is a one-sided social dilemma. Building on the binary trust game, we ask how trust and trustworthiness can evolve in a population where partners are matched randomly and agents sometimes act as trustors and sometimes as trustees. Trustors have the option to costly check a trustee's last action and to condition their behavior on the signal they receive. We show that the resulting population game admits two components of Nash equilibria. Nevertheless, the long-run outcome of an evolutionary social learning process modeled by the best response dynamics is unique. Even if unconditional distrust initially abounds, the trustors' checking option leads trustees to build a reputation for trustworthiness by honoring trust. This invites free-riders among the trustors who save the costs of checking and trust blindly, until it does no longer pay for trustees to behave in a trustworthy manner. This results in cyclical convergence to a mixed equilibrium with behavioral heterogeneity where suspicious checking and blind trusting coexist while unconditional distrust vanishes.

JEL classification: C72; C90

Keywords: Trust Game, Evolutionary Game Theory, Reputation, Best Response Dynamics

*Institute for Analytical Economics, WU Vienna University of Economics and Business (ulrich.berger@wu.ac.at)

1 Introduction

1.1 Trust

Trust is a characteristic that throughout the world permeates a vast variety of human relationships. Friendship and love, family relations, economic and business relations, are all built to some extent on trust. Arrow (1973, 1974) described trust as an "important lubricant of a social system" without which "no market could function". Economists have pointed out that trust is an important part of human capital, through various channels influencing macroeconomic variables like GDP growth, inflation and volume of trade (La Porta et al. (1997), Zak and Knack (2001), Guiso et al. (2006)). The causes and consequences of trust have also been intensively studied in sociology, psychology, and management science.

Here we focus on the question how interpersonal trust can arise in the first place. An interpersonal trust situation is an interdependent relationship between a trustor and a trustee that can be well described by a two-person game. The paradigmatic trust game is the one constructed by Berg et al. (1995), also called the investment game, but this is an infinite game and here we focus on a simpler binary version as it was presented in Dasgupta (1988) and in Kreps (1990).

In a binary trust game, the trustor may or may not place trust. If he does, the trustee has the options to honor or to abuse trust. Compared to the status quo of no trust being placed, the trustor benefits from honored trust but regrets abused trust, while the trustee benefits from honoring trust but even more so from abusing trust. Formally, this defines an extensive form game. The trustor starts the game with the options to trust (T) or to distrust (D). If he distrusts, the game ends. If he trusts, then the trustee can react by honoring (H) or abusing (A) trust. The trustor's preferences over outcomes are given by $(T, H) \succ (D) \succ (T, A)$ and the trustee's preferences are $(T, A) \succ (T, H) \succ (D)$. The unique subgame-perfect Nash equilibrium requires the trustee to abuse any trust placed and therefore the trustor to

distrust. However, this equilibrium is strictly Pareto dominated by the trust-honor outcome. Hence, individual rationality is incompatible with social optimality; the trust game represents a social dilemma. Note, however, that as opposed to the prisoners dilemma the trust dilemma is only one-sided. If the trustee could credibly commit to choosing H, the trustor would happily place trust.

1.2 Trustworthiness

If placing trust is not part of a subgame-perfect Nash equilibrium, then why do we observe so much trust? A simple answer is that many trustees appear to be trustworthy, i.e. they honor trust. This does not only correspond to everyday experience but has been observed in literally hundreds of laboratory experiments (Johnson and Mislin (2011)). Given a positive probability of a stranger to behave in a trustworthy manner, the trustor's optimal decision depends on the stakes and on his risk preferences only and may well be to place trust. But this just raises the more puzzling question of why trustees should be trustworthy at all.

Intuitively, in the absence of explicit incentives like contractually agreed rewards or punishment, being trustworthy may nevertheless pay off if (i) encounters between trustors and trustees are repeated, and if (ii) there is a chance that a trustee's trustworthiness becomes known among trustors. In this case honoring trust may serve as a means to gain a reputation of being trustworthy. If trustors condition their behavior on trustees' reputations, establishing a reputation for trustworthiness may turn out to be advantageous.[1]

[1] Indeed, trustors have a habit of checking the past behavior of their trustees. *Trust, but verify!* is a catchphrase Ronald Reagan famously used, derived from a Russian proverb frequently cited by Vladimir Lenin. The German *Trau, schau, wem!* with the same meaning goes back to the Latin *Fide, sed cui, vide!* Apparently, forming one's beliefs about a trustee's expected behavior by examining his past trustworthiness has stood the test of time.

1.3 Evolution of trust and trustworthiness

The traditional economic approach to model reputation effects in repeated strategic interactions is via the theory of repeated games. Kreps (1990) is an early example where the trust game has been subjected to an analysis in this flavor; for a more recent one see Xie and Lee (2012). While the repeated games approach to reputation has accumulated an impressive amount of literature (Mailath and Samuelson (2006)), it is inherently static and therefore unable to explain why or how different modes of behavior may arise from given initial settings. Moreover, the Folk Theorem typically destroys the hopes of arriving at a unique prediction for equilibrium behavior.

During the last two decades, the evolution-and-learning approach to modeling changes in behaviors within populations of interacting individuals has gained more ground. Pioneered in biology by Maynard Smith (1982), evolutionary game theory later gained popularity in economics as well (Samuelson (1997), Weibull (1997), Fudenberg and Levine (1998), Hofbauer and Sigmund (1998), Gintis (2000), Cressman (2003), Sandholm (2010)). The evolutionary approach typically posits a population of boundedly rational players who are repeatedly and randomly matched to interact in a game. Each player is bound to a strategy for some time, but strategies may be revised every now and then and this makes the proportions of strategies in the population evolve, with successful strategies becoming more frequent over time.

We follow the evolutionary approach here and study a simple model of the evolution of trust and trustworthiness in a single population under the best response dynamics, a traditional learning model for rational but myopic agents. Since agents are randomly matched, trust can only arise if there is the possibility to obtain information about the trustee's behavior. Previous models of this type typically assumed that this information is provided for free to all or a subset of all trustors.[2] However, in many instances such information is not public and must be actively obtained by trustors, which is

[2]See the next section for details on these models.

a costly endeavor. Here we assume that obtaining information on a trustee's previous behavior is costly and that the decision whether or not to "purchase" this information (an act we call *checking*) is a strategic decision of the trustor. This leads to a relatively simple model where individuals prepare strategies for both roles they might find themselves in. In the trustee role these are either honor or abuse, while in the trustor role strategies differ in whether or not they check and, in the checking case, how to react conditional on the information they receive. While the full dynamics of the population state takes place in a large 11-dimensional space, it turns out that it can be reduced in three simple steps to the analysis of a 2-dimensional dynamical system which is straightforward to solve.

We show analytically that in the long run we can expect the state of the population to converge cyclically to an equilibrium state with behavioral heterogeneity: Trustworthy and non-trustworthy trustees coexist with trustors who check their partner's previous behavior and with "blindly trusting" trustors who abstain from checking. Unconditional distrust, however, is bound to disappear in the long run.

1.4 Related literature

We are not the first to study trust and trustworthiness in an evolutionary setting. The amount of published work is still tiny compared to the (closely related) literature on the evolution of cooperation in the prisoner's dilemma, but several rather diverse evolutionary models are scattered throughout the economic, biological and social sciences literature. An early approach can be found in a series of papers by Güth and Kliemt (1994, 2000) and Güth et al. (2000). This work is based on the *indirect* evolutionary approach which assumes that behavior is rational but preferences evolve. The focus is on the conditions under which a trustworthy type (i.e. a type of trustee who prefers to honor trust even in a one-shot setting) can evolve. For more recent studies in this spirit see Ahn and Esarey (2008) or Rabanal and Friedman (2014, 2015).

In the more traditional ("direct") evolutionary approach we employ, preferences are fixed, but strategies evolve. Complex models where trusting need not be a binary choice and various assumptions on the specifics of the transfer of information from trustees to trustors are employed have been analyzed numerically by Bicchieri et al. (2004), Bravo and Tamburino (2008), McNamara et al. (2009), Manapat and Rand (2012), and Manapat et al. (2013). Analytical studies of evolutionary trust models can be found in De Silva and Sigmund (2009), Courtois and Tazdaït (2012), Masuda and Nakamura (2012), and Tarnita (2015). These studies differ in various aspects from our analysis. While they all show that in the long run mutual trust and trustworthiness can be established at least to some degree, the specific mechanisms how and why this happens is sometimes obscured.

Interestingly, the work most closely related to ours turns out to be Andreozzi (2013). It studies the evolution of trust under best response dynamics in a population where players are modeled by finite automata. Upon being matched, two automata play infinitely many rounds of the binary trust game and receive their discounted payoffs, diminished by the complexity costs of the automaton they use. Remarkably, while this approach is very different from ours, it leads to a qualitatively identical dynamical system (apart from Andreozzi (2013) considering two separate populations for the two player roles), suggesting an interesting parallel between our conditional strategies with checking costs and his finite automata with complexity costs.

We propose the present model for the study of the evolution of trust and trustworthiness mainly for its apparent simplicity. The assumption of costly checking as a strategic decision seems natural. Without having to rely on numerical simulations we are able to predict a unique long-run equilibrium which lends itself to straightforward comparative statics analysis. Moreover, the underlying dynamical process leading to selection of this equilibrium and the resulting behavioral heterogeneity is presented in a visually appealing way which can be understood quite intuitively.

2 Model

2.1 The binary trust game

We follow Kreps (1990) here and consider the binary trust game (TG) depicted in Figure 1. Player 1, the *trustor*, can either trust (T) or distrust (D). Distrust ends the game with payoffs 0 for both players. If the trustor trusts, player 2, the *trustee*, can either honor the trust placed in him (H), resulting in rewards $r_1 > 0$ for the trustor and $r_2 > 0$ for the trustee, or he can abuse the trust, capturing the benefit $b > r_2$ for himself while the trustor incurs a loss $-c < 0$. We assume that $r_1 + r_2 > b - c$. The strategic form of this game is given by the payoff bimatrix

$$\begin{array}{c|cc} & H & A \\ \hline T & r_1, r_2 & -c, b \\ \\ D & 0, 0 & 0, 0 \end{array} \tag{1}$$

The efficient outcome is (T, H), but the unique subgame perfect Nash equilibrium is (D, A). This equilibrium is part of a component of partially mixed Nash equilibria $\{(D, q) | \, 0 \leq q \leq \frac{c}{c+r_1}\}$ where q, the probability of the trustee honoring trust, is small enough to render distrust a best response for the trustor.

The symmetrized binary trust game (STG) starts with nature choosing which player will be the trustor and which the trustee. This is a symmetric two-player game with strategies $s_i \in \hat{S} = \{TH, TA, DH, DA\}$, where the first and second component of s_i denotes the action chosen in the role of the trustor and the trustee, respectively. Up to the factor $1/2$, the STG is given by the payoff matrix

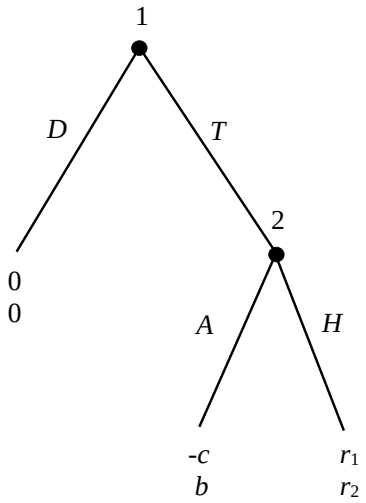

Figure 1: The binary trust game.

$$\begin{array}{c|cccc} & TH & TA & DH & DA \\ \hline TH & r_1+r_2 & r_2-c & r_1 & -c \\ TA & r_1+b & b-c & r_1 & -c \\ DH & r_2 & r_2 & 0 & 0 \\ DA & b & b & 0 & 0 \end{array} \tag{2}$$

Consider now a large population of infinitely-lived individuals each of which is bound to play some pure strategy. Time t is continuous and individuals are repeatedly and randomly matched in pairs to interact in the symmetrized binary trust game. Since we want to integrate the concept of a reputation for trustworthiness, we assume that a trustor may opt to obtain information on the trustee's last action in the trustee-role. We call this behavior *checking*. We presume that there are many more interactions than revision opportunities. Hence, checking amounts to revealing to the trustor whether or not the trustee is trustworthy (plays H in his trustee-role). However, we

assume that checking is also costly, with checking costs $a > 0$ deducted from the trustor's payoff in each interaction where he chooses to check.

Upon playing the trust game with the checking option (TGC), the trustor now has to decide whether or not to check. If he does not check, he has to choose an action T or D. If he checks, he has to choose an action conditional on the signal H or A he receives. The strategy set of a trustor in the TGC is therefore $S_1 = \{NT, ND, TT, TD, DT, DD\}$, where NX is the strategy of not checking and playing action $X \in \{T, D\}$ while the remaining four strategies are checking-strategies denoted by the actions chosen conditional on observing H and A, respectively.

We assume that trustees do not observe whether or not they have been checked, so the trustee's strategy set in the TGC remains $S_2 = \{H, A\}$. In the symmetrized trust game with the checking option (STGC) the players' strategy set is now enlarged to

$$\begin{aligned} S &= \{NTH, NDH, TTH, TDH, DTH, DDH, \\ &\quad NTA, NDA, TTA, TDA, DTA, DDA\}. \end{aligned} \tag{3}$$

Payoffs in the STGC are $\pi(XYZ, X'Y'Z') = \frac{1}{2}[\pi_1(XY, Z') + \pi_2(X'Y', Z)]$, where $XY, X'Y' \in S_1$, $Z, Z' \in S_2$ and the payoff functions π_1 and π_2 are given by the payoff bimatrix

$(\pi_1, \pi_2) =$

	H	A
NT	r_1, r_2	$-c, b$
ND	$0, 0$	$0, 0$
TT	$r_1 - a, r_2$	$-c - a, r_2$
TD	$r_1 - a, r_2$	$-a, 0$
DT	$-a, 0$	$-c - a, b$
DD	$-a, 0$	$-a, 0$

(4)

We denote by Δ_S the 11-dimensional simplex of mixed strategies in the

STGC and extend the payoff function π in the ususal way to mixed strategies $x \in \Delta_S$.

2.2 Best response dynamics

In our player population strategy XYZ earns an expected per-interaction-payoff of $\pi(XYZ, x)$, where $x \in \Delta_S$ is the population state. We posit now a basic version of bounded rationality and assume that strategy updating is guided by the social learning dynamics known as the *best-response dynamics* (BR-dynamics), which was introduced by Gilboa and Matsui (1991), Matsui (1992), and Hofbauer (1995). Under the BR-dynamics, each player is equipped with a Poisson alarm clock. Upon his clock ringing, a player revises his strategy choice by choosing a myopic pure best response to the current population state. This results in the population state $x(t)$ moving along (possibly non-unique) solutions, called best response paths (BR-paths), of the differential inclusion

$$\dot{x}(t) \in B(x(t)) - x(t), \tag{5}$$

where $B(x)$ is the set of (pure or mixed) best responses to state x. As long as the current best response is unique, the BR-path describes a straight line pointing to this current pure best response. If a BR-path converges, the limit is a Nash equilibrium.

For asymmetric two-population games with population states p and q the BR-dynamics reads $(\dot{p}(t), \dot{q}(t)) \in (B_1(q(t)), B_2(p(t))) - (p(t), q(t))$, where B_1 and B_2 are the respective best response correspondences. For example, in the TG all BR-paths outside of the equilibrium component converge to the subgame perfect equilibrium, as shown in Figure 2.

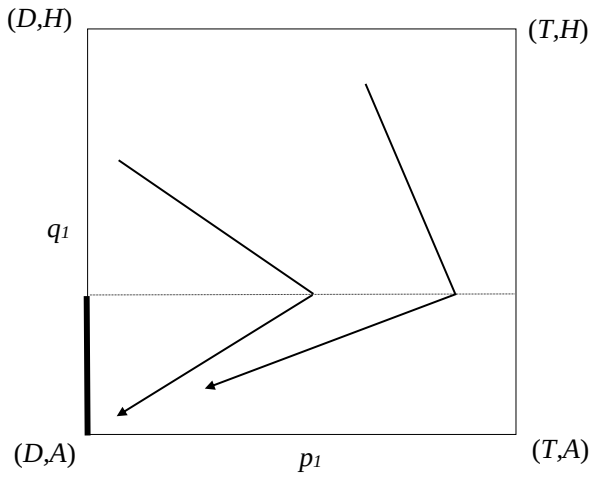

Figure 2: In the TG, all BR-paths outside the equilibrium component converge to (D, A).

3 Analysis

Now we set out to find the long-run behavior of the population state in the STGC under the BR-dynamics. Though the state space is 11-dimensional, the task is greatly simplified by the following three observations, which we will use to sequentially reduce the state space we have to analyze.

3.1 Observation 1

Under BR-dynamics updating players never switch to strictly dominated strategies, hence such strategies are eliminated quickly. It thus suffices to study the BR-dynamics in the reduced STGC after eliminating strictly dominated strategies. From the payoff bimatrix (4) it is clear that the trustor's strategies TT, DT, DD are strictly dominated. Intuitively, it makes no sense to pay the checking costs and then *not* to react optimally to the signal obtained. It follows that in the STGC the six strategies TTH, DTH, DDH,

TTA, DTA, DDA are strictly dominated as well. Eliminating those we are left with the reduced STGC (rSTGC) comprising the remaining six strategies NTH, NDH, TDH, NTA, NDA, TDA. This reduces the dimension of the state space we are dealing with from eleven to five.

The STGC is only interesting if checking is not prohibitively costly, i.e. if the trustor's strategy TD is not dominated by a mixture of NT and ND. A quick calculation shows that for this we have to assume

$$a < \frac{r_1 c}{r_1 + c}, \tag{6}$$

which we will do henceforth.

3.2 Observation 2

Consider the projection from S to $S_1 \times S_2$ which separates a strategy XYZ into the corresponding pair of strategies (XY, Z), and its extension to Δ_S. From Berger (2002), this projection respects the best response structure of the game, i.e. it maps BR-paths of the rSTGC to BR-paths of the corresponding reduced TGC (rTGC) given by the payoff functions[3]

$$(\pi_1, \pi_2) = \begin{array}{c|cc} & H & A \\ \hline NT & r_1, r_2 & -c, b \\ ND & 0, 0 & 0, 0 \\ TD & r_1 - a, r_2 & -a, 0 \end{array} \tag{7}$$

From the long-run behavior of the population state in the rTGC we can then infer the long-run behavior of the population state in the rSTGC. This allows us to further reduce the dimension of the state space we are working in from five to three.

[3] With a little abuse of notation we stick to denoting those payoff functions by π_1 and π_2.

3.3 Observation 3

The rTGC is a two-person game where one of the players has only two pure strategies. For this class of games Berger (2005) showed that all BR-paths converge to the set of Nash equilibria. Moreover, a suitable projection allows one to analyze the global dynamics in these games in two dimensions. The projection is chosen in such a way that the hyperplane of indifference of the player with two strategies (the trustee in case of the rTGC) is projected to a horizontal line. This projection maps the state space $\Delta_3 \times \Delta_2$ of the rTGC and its partition into different best response regions to a rectangle partitioned into different rectangles in the plane. For the rTGC with population states $(p, q) \in \Delta_3 \times \Delta_2$ the projection is given by the map

$$P(p, q) = (q_2, \pi_2(p, H) - \pi_2(p, A)). \tag{8}$$

We call $P(p, q)$ the *induced population state*. As explained in more detail in Berger (2005), P maps BR-paths in the rTGC to so-called *induced paths* in the plane. The behavior of induced paths in the plane is easy to study and allows one to obtain the behavior of BR-paths in the rTGC and, by Observation 2, in the rSTGC and therefore in the original STGC.

3.4 Combining the Observations

The combination of the three observations listed above allows us to reduce the study of BR-paths in the 11-dimensional state space of the STGC to the study of induced paths in the 2-dimensional plane. Since the projection maps are linear, piecewise linear BR-paths pointing to pure strategy pairs in the rTGC are mapped to piecewise linear induced paths pointing to points on the boundary of the induced state space (the rectangle) in the plane. The remaining analysis is a simple exercise in planar geometry and the result is shown in Figure 3.

By construction of P, trustees switch to honoring if the induced population

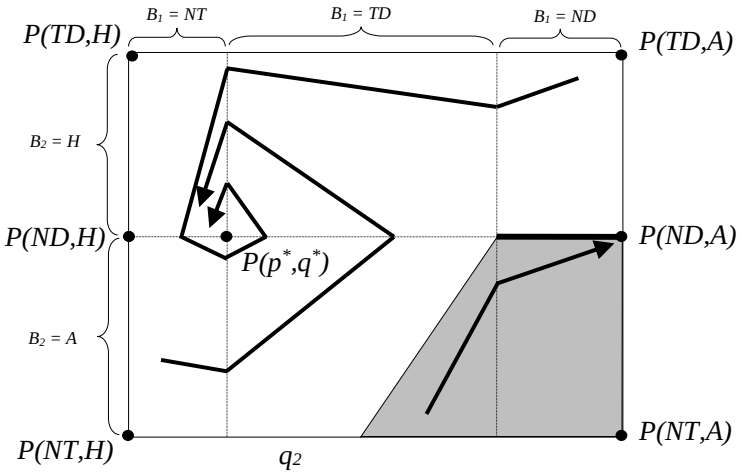

Figure 3: Induced paths $(q_2(t), \pi_2(p(t), H) - \pi_2(p(t), A))$ in $P(\Delta_3 \times \Delta_2)$ of the rTGC. The middle line is the horizontal axis.

state is above the horizontal axis and they switch to abusing if it is below the horizontal axis. Induced paths therefore move to the left above and to the right below the horizontal axis.

If q_2 is large, i.e. if most trustees abuse, trustors switch to blind distrust (ND), since neither blind trust nor buying information pays off. In the rightmost vertical sector, therefore, induced paths point to one of the boundary points of the horizontal axis. This is the case for $q_2 > \frac{r_1 - a}{r_1}$. If q_2 is in an intermediary range, $\frac{a}{c} < q_2 < \frac{r_1 - a}{r_1}$, the checking strategy TD becomes optimal for trustors. In the middle vertical sector induced paths therefore point to one of the top vertices of the rectangle. Finally, if q_2 is small enough, $q_2 < \frac{a}{c}$, it does no longer pay to check and trustors turn to trusting blindly (NT). In the leftmost vertical sector induced paths therefore point to one of the bottom vertices of the rectangle.

The rTGC has a mixed Nash equilibrium (p^*, q^*) at a point on the boundary face of the state space where blind distrust is unused and where trustors

are just indifferent between blind trust and checking (both strictly superior to blind distrust), while trustees are just indifferent between honoring and abusing trust. This equilibrium can be calculated to be given by

$$p^* = \left(\frac{r_2}{b}, \frac{b-r_2}{b}, 0\right), \tag{9}$$

$$q^* = \left(\frac{c-a}{c}, \frac{a}{c}\right). \tag{10}$$

BR-paths in a neighborhood of this equilibrium spiral inwards and converge cyclically to the equilibrium, as is well known from Brown's (1951) early analysis of the mathematically equivalent continuous-time fictitious play process in zero-sum games.[4] The induced equilibrium $P(p^*, q^*)$ and the induced paths spiraling into it can be seen in Figure 3.

Apart from this isolated equilibrium the rTGC also admits a 1-dimensional component of Nash equilibria where trustors blindly distrust and q_2 is large enough to yield ND the trustors' best response. Payoffs are zero for all players in this equilibrium component. The induced component is visible as the bold line segment on the horizontal axis in Figure 3.

From Figure 3 it is clear that all induced paths outside of the induced equilibrium component converge either cyclically to the induced mixed equilibrium $P(p^*, q^*)$ or straight to the induced equilibrium $P(ND, A)$. The latter's basin of attraction is marked as the grey region in the rectangle. For the rTGC this means that all BR-paths outside the equilibrium component converge either to the equilibrium (p^*, q^*) or to (ND, A). BR-paths starting in the equilibrium component are non-unique. They may stay in the component forever or move to the left end of the component and then exit towards (TD, H) and finally converge to (p^*, q^*).

Note, however, that while (p^*, q^*) is locally asymptotically stable, the equi-

[4] Cyclic 2×2-games are strategically equivalent to a zero-sum game, see Hofbauer and Sigmund (1998). Geometric proofs for cyclic convergence can be found in Rosenmüller (1971), Berger (2002) and Berger (2012).

librium component is *unstable*. If an arbitrarily small fraction of trustors starts to check, honoring suddenly beats abusing and the population state leaves the component's neighborhood and converges to (p^*, q^*). While this is extraneous to the model, it implies that under arbitrarily small noise we should expect the population to end up at (p^*, q^*) in the long run from any initial condition.

The final step in our analysis is to move back from the rTGC to the rSTGC. This can be done as explained in Berger (2002). The equilibrium (p^*, q^*) in the rTGC corresponds to a continuum of equilibria in the rSTGC. However, there is only a single equilibrium in this continuum which attracts the BR-paths in the rSTGC. This is called the Wright equilibrium, since it lies at the intersection of the equilibrium continuum with the so-called Wright manifold (see also Cressman (2003) for details on the role of the Wright manifold in symmetrized games). In this equilibrium the frequency of the rSTGC's pure strategy XYZ is given by the product of the frequencies of the respective rTGC's pure strategies XY and Z in (p^*, q^*), which can thus be written as

$$\begin{aligned} x^* &= \\ &= \frac{r_2(c-a)}{bc}NTH + \frac{r_2 a}{bc}NTA + \frac{(b-r_2)(c-a)}{bc}TDH + \frac{(b-r_2)a}{bc}TDA. \end{aligned} \tag{11}$$

Of course (with a little abuse of notation) this is also the attracting equilibrium in the original STGC.

A more illustrative way to write down the trustors' and the trustees' long-run strategies (9) and (10) is perhaps

$$p^* = \frac{r_2}{b}NT + \frac{b-r_2}{b}TD, \quad q^* = \frac{c-a}{c}H + \frac{a}{c}A. \tag{12}$$

Long-run equilibrium payoffs in the STGC are $\frac{1}{2}\left[r_2 + r_1\frac{c-a}{c} - a\right] > 0$ and the frequency of efficient interactions (blind trust and honor) is $\frac{r_2(c-a)}{bc}$. The social dilemma is therefore partially remedied, depending on parameter

values.

4 Discussion

Interestingly, from the equilibrium frequencies in (11) and (12) it can be seen that r_1 has no influence whatsoever on the levels of trust or trustworthiness.[5] Intuitively, this is because trustors are rewarded whenever they meet a trustworthy trustee, independently of whether they check or trust blindly. Thus they only have to trade off the costs of checking and the costs of being abused in their indifference condition.

It is also noteworthy that the frequencies show the usual counterintuitive reaction to payoff changes in mixed equilibria. While one could superficially expect that the demand for checking goes up if the price a of checking decreases, this is not the case. The equilibrium frequency of the checking option remains unchanged. Instead, honoring of trust increases among trustees. The reason, of course, is that in a mixed equilibrium the frequencies of a player's strategies are determined by indifference of the *other* player. Indeed, if a goes to zero, abusing vanishes among trustees, while the ratio between blind trust and checking remains constant. In Figure 3 the induced equilibrium component of distrust on the right shrinks to a point at $P(ND, A)$ while the induced isolated equilibrium moves to the left and converges to the boundary along the horizontal axis. Note, however, that there is a discontinuity at the limit $a = 0$. If we let checking be costless, the trustor's checking strategy TD becomes weakly dominant and consequently the upper left vertex in Figure 3, corresponding to the STGC-strategy TDH, attracts all interior best response paths. In the limit, all interactions are efficient.

[5] Provided r_1 is not too low, i.e. as long as inequality (6) is guaranteed.

References

Ahn, TK and Justin Esarey (2008), "A dynamic model of generalized social trust." *Journal of Theoretical Politics*, 20, 151–180.

Andreozzi, Luciano (2013), "Evolutionary stability in repeated extensive games played by finite automata." *Games and Economic Behavior*, 79, 67–74.

Arrow, Kenneth J (1973), "Information and economic behavior." Technical report, DTIC Document.

Arrow, Kenneth J (1974), *The Limits of Organization*. New York: Norton.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995), "Trust, reciprocity, and social history." *Games and Economic Behavior*, 10, 122–142.

Berger, Ulrich (2002), "Best response dynamics for role games." *International Journal of Game Theory*, 30, 527–538.

Berger, Ulrich (2005), "Fictitious play in 2×n games." *Journal of Economic Theory*, 120, 139–154.

Berger, Ulrich (2012), "Non-algebraic convergence proofs for continuous-time fictitious play." *Dynamic Games and Applications*, 2, 4–17.

Bicchieri, Cristina, John Duffy, and Gil Tolle (2004), "Trust among strangers." *Philosophy of Science*, 71, 286–319.

Bravo, Giangiacomo and Lucia Tamburino (2008), "The evolution of trust in non-simultaneous exchange situations." *Rationality and Society*, 20, 85–113.

Courtois, Pierre and Tarik Tazdaït (2012), "Learning to trust strangers: an evolutionary perspective." *Journal of Evolutionary Economics*, 22, 367–383.

Cressman, Ross (2003), *Evolutionary Dynamics and Extensive Form Games*. MIT Press.

Dasgupta, Partha (1988), "Trust as a commodity." In *Trust: Making and Breaking Cooperative Relations*, 49–72, Basil Blackwell. Oxford.

De Silva, Hannelore and Karl Sigmund (2009), "Public good games with incentives: the role of reputation." In *Games, Groups, and the Global Good*, 85–103, Springer.

Fudenberg, Drew and David K Levine (1998), *The Theory of Learning in Games*. MIT Press.

Gilboa, Itzhak and Akihiko Matsui (1991), "Social stability and equilibrium." *Econometrica*, 59, 859–867.

Gintis, Herbert (2000), *Game Theory Evolving*. Princeton University Press.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (2006), "Does culture affect economic outcomes?" *Journal of Economic Perspectives*, 20, 23–48.

Güth, Werner and Hartmut Kliemt (1994), "Competition or co-operation: On the evolutionary economics of trust, exploitation and moral attitudes." *Metroeconomica*, 45, 155–187.

Güth, Werner and Hartmut Kliemt (2000), "Evolutionarily stable co-operative commitments." *Theory and Decision*, 49, 197–222.

Güth, Werner, Hartmut Kliemt, and Bezalel Peleg (2000), "Co-evolution of preferences and information in simple games of trust." *German Economic Review*, 1, 83–110.

Hofbauer, Josef (1995), "Stability for the best response dynamics." Technical report, University of Vienna.

Hofbauer, Josef and Karl Sigmund (1998), *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Johnson, Noel D. and Alexandra A. Mislin (2011), "Trust games: A meta-analysis." *Journal of Economic Psychology*, 32, 865–889.

Kreps, David M (1990), "Corporate culture and economic theory." In *Perspectives on Positive Political Economy*, 90–143, Cambridge University Press.

La Porta, Rafael, Florencio Lopez-de Silanes, Andrei Shleifer, and Robert W. Vishny (1997), "Trust in large organizations." *American Economic Review*, 87, 333–338.

Mailath, George J and Larry Samuelson (2006), *Repeated Games and Reputations*. Oxford: Oxford University Press.

Manapat, Michael L, Martin A Nowak, and David G Rand (2013), "Information, irrationality, and the evolution of trust." *Journal of Economic Behavior & Organization*, 90, 57–75.

Manapat, Michael L and David G Rand (2012), "Delayed and inconsistent information and the evolution of trust." *Dynamic Games and Applications*, 2, 401–410.

Masuda, Naoki and Mitsuhiro Nakamura (2012), "Coevolution of trustful buyers and cooperative sellers in the trust game." *PLoS ONE*, 7, e44169.

Matsui, Akihiko (1992), "Best response dynamics and socially stable strategies." *Journal of Economic Theory*, 57, 343–362.

Maynard Smith, John (1982), *Evolution and the Theory of Games*. Cambridge University Press.

McNamara, John M, Philip A Stephens, Sasha RX Dall, and Alasdair I Houston (2009), "Evolution of trust and trustworthiness: social awareness favours personality differences." *Proceedings of the Royal Society of London B*, 276, 605–613.

Rabanal, Jean Paul and Daniel Friedman (2014), "Incomplete information, dynamic stability and the evolution of preferences: Two examples." *Dynamic Games and Applications*, 4, 448–467.

Rabanal, Jean Paul and Daniel Friedman (2015), "How moral codes evolve in a trust game." *Games*, 6, 150–160.

Rosenmüller, Joachim (1971), "Über Periodizitätseigenschaften spieltheoretischer Lernprozesse." *Probability Theory and Related Fields*, 17, 259–308.

Samuelson, Larry (1997), *Evolutionary Games and Equilibrium Selection*. MIT Press.

Sandholm, William H (2010), *Population Games and Evolutionary Dynamics*. MIT Press.

Tarnita, Corina E. (2015), "Fairness and trust in structured populations." *Games*, 6, 214–230.

Weibull, Jörgen W (1997), *Evolutionary Game Theory*. MIT press.

Xie, Huan and Yong-Ju Lee (2012), "Social norms and trust among strangers." *Games and Economic Behavior*, 76, 548–555.

Zak, Paul J and Stephen Knack (2001), "Trust and growth." *Economic Journal*, 111, 295–321.