Delayed and Inconsistent Information and the Evolution of Trust

Michael L. Manapat · David G. Rand

Published online: 1 September 2012 © Springer Science+Business Media, LLC 2012

Abstract Trust is essential in virtually all economic interactions. In this paper, we build on previous work analyzing the "trust game" from the perspective of evolutionary dynamics. There are two players: an "investor" and a "trustee." The investor begins with one monetary unit and can choose to transfer it to the trustee. The transfer, if made, is multiplied by a factor b > 1 representing the gains that arise from cooperation. The trustee can then return a fraction of his choosing to the investor. Previous work has shown that adding information to this game can lead to trusting and trustworthy behavior. But in those models, information spreads instantaneously and investors never face conflicting information. Here, we relax both of these assumptions. We introduce delays in information propagation so that an investor may still be acting on old information after a trustee has changed his behavior. And we give investors "memories" and thereby allow for the possibility that they might face conflicting information about trustees. In both cases, we find that the trust and trustworthiness induced by information is robust to delays and conflicts. Even if it takes time for information to spread, and even if investors sometimes deem information to be unreliable, the benefits of trust are realized with just moderate levels of information about trustees. We conclude that information (or "reputation") is a robust explanation for the trust and trustworthiness observed among humans.

Keywords Trust game · Reputation · Evolutionary game theory · Evolutionary dynamics

1 Introduction

Trust is an essential component of social and commercial interactions. When individuals make purchases online, they trust that sellers have not misrepresented the characteristics of

M.L. Manapat (🖂)

Program for Evolutionary Dynamics and School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA e-mail: manapat@fas.harvard.edu

the goods and that they will actually ship the items. When pension funds, university endowments, and retail investors transfer money to investment management organizations, they trust that the managers will fulfill their fiduciary obligations and handle the funds properly. (A number of recent scandals have shown that this trust can be misplaced, with devastating consequences.) In general, whenever a principal employs an agent to perform some task, the principal must trust the agent to act in a manner consistent with the principal's interests. Trust has important implications for the efficiency of society [8, 9, 13, 15, 24]. As Kenneth Arrow put it, trust is "an important lubricant of a social system" [1].

In this paper, we study how information affects the evolution of trust by analyzing the trust game [2–7, 11, 12, 14, 16, 17]. There are two players, an "investor" and a "trustee." The investor begins with one monetary unit. Her strategy is given by the probability p_0 that she transfers the unit to the trustee. If the investor does make the transfer, the trustee receives b > 1. The trustee's strategy is given by the fraction r of what he receives that he returns to the investor. The investor's expected payoff is $1 - p_0 + bp_0r$ and the trustee's expected payoff is $bp_0(1 - r)$. The parameter p_0 is a measure of how "trusting" the investor is. The parameter r is a measure of how "trustworthy" the trustee is. A "trusting" investor transfers her endowment with high probability. A "trustworthy" trustee returns a relatively large fraction of what he receives. The trust game allows us to study the evolution of trust game-theoretically. Player 1 is the investor or the principal. Player 2 is the trustee or the agent. The multiplicative factor b determines how beneficial the game can be to the players. In some sense, b is the value that the trustee provides to the investor, i.e., the value of agency.

In a one-shot anonymous trust game, the unique subgame perfect Nash equilibrium is $p_0 = 0$, r = 0. A rational, self-interested trustee should not return anything to the investor. Hence, a rational, self-interested investor should not make the transfer. This game-theoretic analysis does a poor job of predicting actual levels of trust and trustworthiness, however. Numerous behavioral experiments have shown high levels of both trust and trustworthiness in one-shot anonymous interactions [11]. Why are people trusting and trustworthy, even when it is "irrational" to be so?

Previous work has shown that the presence of "information" (or "reputation") is one possible answer to this question. McNamara et al. [19] considered a binary trust game, studied initially by Guth and Kliemt [10], in which investors could obtain, at a cost, information about trustees' past behavior. In the binary game, the investor decides whether or not to make the transfer. If she does not, both players receive a payoff of *s*. If she does, then the trustee must decide whether to keep everything he receives or to return a fair amount to the investor. In the former case, the investor's payoff is 0 and the trustee's payoff is 1. In the latter case, both players receive a payoff of *r*. The payoff values satisfy the inequalities 0 < s < r < 1.

McNamara et al. [19] allowed investors to condition their decision on the past behavior of trustees as follows. By paying a cost, an investor can learn what a given trustee did in n past interactions ($n \ge 0$ is a fixed parameter of the model). Each investor has a strategy parameter k, $0 \le k \le n + 1$. If k = 0, the investor always makes the transfer (and never incurs the cost). If k = n + 1, the investor never makes the transfer (and never incurs the cost). If $1 \le k \le n$, then the investor pays the cost to learn what the trustee did n times in the past and makes the transfer if and only if the trustee "cooperated" at least k out of the ntimes.

McNamara et al. found that if $n \ge 2$ (or if n = 1 and the mutation rate is sufficiently high to maintain a high degree of diversity among trustees), then both trusting behavior (in the form of investor k's not equal to n + 1) and trustworthy behavior (in the form of positive probabilities of choosing the "cooperative" action) can arise. In Sect. 4, we will consider a

Dyn Games Appl (2012) 2:401-410

model that is similar—though not precisely a generalization—of theirs in which investors base their decisions on the consistency of the information they have about trustees' past behavior.

In [19], investors always have access to information if they are willing to pay the cost. Manapat, Nowak, and Rand [17] instead considered a scenario in which there is a fixed probability of having information about the trustee but no guarantee that information is available in any given interaction. In their continuous trust game with information, investors know with probability q the return fraction of the trustee before having to decide whether to make the transfer. They found that trust—i.e., willingness on the part of investors to make the transfer when information is *not* available—and trustworthiness arise as long as $q \ge 1/b$. Hence, it is not necessary for information to be universally available for cooperative behavior to be favored by selection. Their model also made explicit how interactions when information is available shape behavior when information is not available, helping to explain the trust and trustworthiness humans display in one-shot anonymous interactions (when information cannot be obtained at any cost).

In both models, the information about investors, when available, is perfectly reliable. In particular, the models omit the fact that information about trustee behavior takes time to propagate after a trustee changes his strategy. Here, we relax this assumption and find that the trust and trustworthiness that arise when investors have information about trustees are robust to delays in the propagation of that information. We describe the basic setup in Sect. 2, introduce delays in information propagation in Sect. 3, and conflicting information in Sect. 4, and conclude in Sect. 5.

2 The Model

Manapat, Nowak, and Rand [17] introduced information into the trust game with the following model. There is a well-mixed population of size N, split evenly between investors and trustees. Each investor has a strategy p_0 , the probability that she makes the transfer when she knows nothing about the trustee. We call p_0 the investor's "trust" as it is a measure of her willingness to make the transfer when she knows nothing about the trustee. Each trustee has a strategy r, the fraction of what he receives that is returned to the investor. We call r the trustee's "return" as it is the fraction of the transfer received that is returned to the investor. In an anonymous one-shot trust game, the investor's expected payoff is $1 - p_0 + bp_0r$ and the trustee's expected payoff is $bp_0(1 - r)$.

Information is incorporated into the game as follows. Fix a q such that $0 \le q \le 1$. When an investor and a trustee meet, there are two possible outcomes. With probability q, the investor knows the trustee's return fraction r before the transaction begins and can condition her behavior on that r. (This simplified model captures the essence of how information might spread in a population via gossip. See [17] for details.) We assume that investors behave in a payoff-maximizing manner: they make the transfer if and only if r > 1/b. Thus, investors act to capture any profit, however small. With probability 1 - q, on the other hand, the investor knows nothing about the trustee and makes the transfer with probability p_0 . We call q the (probability of) information as it is a measure of how well information spreads among investors.

The population evolves as follows. Initially, each p_0 and r is selected uniformly at random from the interval [0, 1]. In each update round, a large number of games are played between randomly selected investors and trustees. Every individual thereby earns a payoff. Two investors, call them A and B, are then selected uniformly at random. Let $\bar{\pi}_A$ be the average per-game payoff of A, $\bar{\pi}_B$ the average per-game payoff of B, and

$$\rho = \frac{1}{1 + e^{-\beta(\bar{\pi}_A - \bar{\pi}_B)}}.$$
(1)

B is replaced by a copy of *A* with probability $\rho(1 - \mu)$ and by a random mutant (with a strategy chosen uniformly at random from [0, 1]) with probability μ . With probability $(1 - \rho)(1 - \mu)$, all the investors retain their original strategies. This process is then repeated for trustees. The parameter μ is the mutation rate. The parameter β is the intensity of selection. The larger β is, the more likely it is that a player will imitate the strategy of someone doing better (and not imitate the strategy of someone doing worse). The "stationary" values of the investor trust p_0 and trustee return *r* are found by averaging over the last 80 % of rounds of this pairwise comparison process [23]. We can view the evolution of p_0 and *r* as the result of genetic evolution (shaping instinctual behavior) or cultural evolution (as investors and trustees learn from each other).

In our model, the evolutionary dynamics converge to the Nash equilibrium (when a unique one exists) described in the following theorem, proven in more generality in [17].

Theorem 1 *The following are all the pure Nash equilibria of the trust game with information:*

 $- p_0 = 0$ and $r \le 1/b$ when q = 0,

 $- p_0 \in [0, 1] and r = 1/b + \epsilon when q = 1,$

- $p_0 = 1$ and $r = 1/b + \epsilon$ when 1/b ≤ q < 1.

(The quantity ϵ corresponds to the smallest possible transferrable amount. Thus, $r = 1/b + \epsilon$ means that the trustee returns the smallest amount strictly greater than 1/b.)

When 0 < q < 1/b, there are no (pure) equilibria. The stationary values of the evolutionary dynamics are then averages over cycles as investors oscillate between $p_0 = 1$ and $p_0 = 0$ while trustees oscillate between $r = 1/b + \epsilon$ and r = 0. When q is just slightly larger than 0, these cycles spend more time in the neighborhood of $p_0 = 0$ and r = 0, so the average p_0 and r are small. When q is just slightly less than 1/b, these cycles spend more time in the neighborhood of $p_0 = 1$ and $r = 1/b + \epsilon$, so the average p_0 and r are large. When $0 \ll q \ll 1/b$, the average strategies assume intermediate values [17].

The model of information diffusion described above is very simple. It assumes that once a trustee changes his strategy, an investor has the same probability q of knowing the trustee's new return fraction r at the beginning of the next round. It also assumes that investors never receive incorrect or conflicting information. In this paper, we explore what happens when these assumptions are discarded.

3 Delayed Information

We begin with the following scenario. Suppose that information about a trustee's return fraction r takes time to propagate through the population of investors. Sometimes, when an investor "knows" the trustee's r before the interaction, the value of r is an old one. As the trustee plays more and more games with the new r, the probability that an investor is informed of the new (correct) value increases. Let us make this idea precise.

Two numbers are associated to each trustee. The first is his strategy parameter r. This is the fraction of what he receives that is actually returned. The second is the what investors

Author's personal copy

Dyn Games Appl (2012) 2:401-410

Fig. 1 The average investor trust p_0 (a) and trustee return Γ (b) as functions of the information update probability U. Trust and trustworthiness are relatively robust to delays. The no-delay equilibrium (U = 1) is maintained until $U \approx 10^{-3}$, at which point p_0 and r begin to fall quickly. There are thus two "regimes," and intermediate levels of trust and trustworthiness are not observed generically. When q = 0.5 and q = 0.75, the strategy parameters stabilize, but when q = 0.25, the values are averages over cycles [17]. We use the following parameters: the multiplier b = 3, the total population size N = 100, the mutation rate $\mu = 0.01$, and the selection intensity = 20. Results are averaged over 25 simulation runs, each run consisting of 100,000 update rounds and each round consisting of 1,000 games between randomly chosen investors and trustees



perceive his return fraction to be (when they think they know it), which we denote by r'. In the fraction q of cases in which the investor knows the trustee's return fraction, it is r' on which the investor bases her decision.

Initially, r = r' for every trustee. When a given trustee updates his strategy, r is updated immediately but r' is not. That is, his actual behavior changes immediately, but his reputation among investors does not. When the trustee subsequently meets an investor in an interaction with information, the investor will know the out-of-date r'. But each time an investor makes the transfer to the trustee (and thereby learns his return fraction q), r' is updated to reflect the new r with probability u. We call u the information update probability. Thus, the number of times a trustee receives the transfer from investors before the new r is correctly reported is geometrically distributed with parameter u. The basic model with information in Sect. 2 corresponds to the case in which u = 1. We note that trustees' return fractions r evolve via the evolutionary process described in Sect. 2, but that their perceived return fractions r' do not evolve—they are simply updated to reflect r after a geometrically distributed "waiting time."

Figure 1(a) shows the average investor trust p_0 as a function of the information update probability u for various values of the information level q. For all but the smallest values of u, the level of trust is the same as it is when u = 1. Figure 1(b) is the analogous plot for the average trustee return r. Again the u = 1 behavior persists even when $u \ll 1$. The trust and trustworthiness induced by the spread of information are thus robust to delays in information propagation.

In Fig. 1, the population size N is 100 and 1,000 games are played per round. This means that each trustee plays on average 20 games per round and 2,000 games between strategy updates. When q = 0.5, investors almost always make the transfer to the trustee (in one simulation run, they did so 95 % of the time over the first 10,000 rounds and 88 % of the

time over all 100,000 rounds). And when $u = 10^{-3}$, the investors begin using the correct information about a particular trustee after 1,000 games on average (under the assumption that the transfer is almost always made). We can thus interpret our results intuitively as follows: investors must use the up-to-date information about trustees in slightly more than half of the games (with information) for trust and trustworthiness to arise and persist. Put another way, we can split the time between strategy updates for a trustee into two parts. During the first part, the trustee's payoff is determined by his interaction with investors who are operating on incorrect information. During the second part, the trustee's payoff is determined by his interaction. Trust and trustworthiness require that the latter part influences the trustee's payoff more than the former part.

Thus far, we have assumed that all investors have the same perception, r, of a given trustee's return fraction. Suppose now that the correct information may reach investors at different times but that the distribution of waiting times is the same for all investors. We formalize this as follows. Investor i perceives the trustee's return fraction to be r_i . Initially, $r_i = r$ for all i. But once the trustee changes his strategy, the r_i s will be stale. Subsequently, every time some investor makes the transfer to the trustee, each r_i is updated to reflect the true r with probability U. Put another way, when any investor makes the transfer to the trustee in question, r_i is updated to reflect the true r with probability U for every i (independently), not just for the i of the investor who made the transfer.

Figures 2(a) and 2(b) are the analogues of 1(a) and 1(b) when updates occur for investors independently. The behavior is qualitatively similar to the simultaneous updates case, but the critical u is about an order of magnitude larger, i.e., trust and trustworthiness are slightly less robust to delays in information. We can reason about this difference as follows. For trust (and trustworthiness) to evolve, roughly all investors need to use correct information in at least a fraction 1/b of interactions. When the update of r to reflect the true r occurs for all investors simultaneously, all investors base their decisions on correct information after a waiting time—i.e., the number of times investors make the transfer to the trustee—that is geometrically distributed with parameter u. On the other hand, when each investor's knowledge is updated independently (but with the same distribution), the waiting time until all the investors are acting on the correct information is roughly N/2 (= the number of investors) times larger. In Figs. 1 and 2, N = 100 and so N/2 = 50, and indeed the critical update probabilities in Fig. 2 are roughly 50 times larger than in Fig. 1.

4 Inconsistent Information

In this section, we consider another modification of Sect. 2's basic model. This time, we allow investors to remember a (partial) history of information they receive about each trustee. When the information they remember is consistent, they act in the payoff-maximizing manner under the assumption that the information is correct. When the information is inconsistent, they assume they know nothing about the trustee and therefore make the transfer with probability p_0 .

We make this precise as follows. Each investor has a queue of maximum length L associated with each trustee. When an investor meets a trustee and is given information about him, that information (the trustee's return fraction r) is added to the investor's queue for the trustee. If the queue already has length L, the oldest reported r is discarded. The investor then examines the entries in the queue. If they are all equal, then the investor takes that r to be the true one and makes the transfer if and only if r > 1/b. If they are not all equal,

406

Dyn Games Appl (2012) 2:401–410

Fig. 2 The average investor trust p_0 (**a**) and trustee return r (**b**) as a function of the information update probability U when updates occur independently for investors. Again trust and trustworthiness are robust to delays, but the critical probability U at which they begin to breakdown is somewhat higher than it is when updates occur for all investors simultaneously. When q = 0.5 and q = 0.75, the strategy parameters stabilize, but when q = 0.25, the values are averages over cycles [17]. We use the following parameters: the multiplier b = 3, the total population size N = 100, the mutation rate $\mu = 0.01$, and the selection intensity $\beta = 20$. Results are averaged over 25 simulation runs, each run consisting of 100,000 update rounds and each round consisting of 1,000 games between randomly chosen investors and trustees



the investor makes the transfer with probability p_0 : she behaves as if she knows nothing about the trustee (taking the inconsistency to mean that the information is unreliable). The longer L is, the less "valuable" information is as there is a higher chance that the queue contains conflicting information. When the investor only remembers the last thing she has heard about each trustee, L = 1, we return to the model of Sect. 2.

Figures 3(a) and 3(b) show how the average trust p_0 and return r vary with the queue length L for various information levels q. Trust and trustworthiness are maintained even for relatively long queues. But once p_0 and r start declining, they both reach their minimum values quickly. As in the case of delays in information propagation, there are two regimes: a high trust and trustworthiness one, and a low trust and trustworthiness one. The transition between them occurs quickly. Unsurprisingly, the larger q is, the more robust the former regime is to increases in queue length: when q is large, inconsistent information is flushed out of the queue more quickly.

Investors in this model are conservative—they act as if there is no information when there is even the slightest inconsistency in what they "remember" about a trustee. One can imagine a situation in which investors are less stringent. For example, they may act as if the trustee's return fraction is a particular r if the majority of entries in their queues equal that r. As we have formulated it here, this means that investors will often be conditioning their behavior on incorrect information: After a trustee updates his strategy, it takes L/2information-based interactions before the true r dominates the queue (assuming no subsequent updates to the trustee's strategy). Hence, this situation is similar to the one examined in the last section in which investors temporarily act on incorrect information after a trustee changes his strategy. Further simulations show that trust and trustworthiness are still robust, though the breakdown occurs at smaller values of L with the "majority" rule than with the "unanimous" rule.

407

Author's personal copy

Queue length, L



5 Conclusion

In this paper, we have introduced three extensions to the evolutionary model of the trust game with information [17]. In the first two of these extensions, investors might possess out-dated information about trustees. The waiting time until the information update is geometrically distributed, and updates can occur for all investors simultaneously or for each investor independently. In the third extension, investors have "memories" and can retain conflicting information about trustees. For all three of these models, we found that information about trustees robustly promotes trusting and trustworthy behavior. Even when updates occur with low probability, and even when investors have memories that can retain conflicting information from long in the past, the trusting and trustworthy equilibrium is maintained for reasonably long delays and memories.

(b)

These results provide further evidence for the hypothesis that information (or "reputation") effects may have contributed to the evolution of trust and trustworthiness in one-shot anonymous interactions. Previous work [17, 19] showed that information could lead to these "desirable" outcomes, and our work here shows that these information-based mechanisms are robust. But further work is needed to explore the conflicts that investors face when spreading information. For example, the model described above and introduced in [17] is based on the following concrete "diffusion" mechanism. After each interaction in which an investor makes the transfer to a trustee (and thereby learns the trustee's return fraction r), the investor informs a fraction q' of the investor population of the trustee's r. As time passes, more and more investors know the r of any given trustee. In [17], it is shown that q' is roughly analogous to q, the probability of knowing the trustee's r in any given interaction. But why do investors tell other investors about their experiences with trustees?

Dyn Games Appl (2012) 2:401-410

409

After all, investors are competing against one another, so spreading information—even if it makes trustees trustworthy—is not an unqualified good. Further work should be done to determine if evolution does indeed select for the spread of information, something that has been imposed exogenously in this paper and other related work [17, 19]. It would also be of interest, particularly in the context of trust in online transactions, to study the effects on trust of allowing agents to buy and sell reputations on a market [20].

We have been studying the effects of information in the trust game assuming that selection is relatively strong—the intensity of selection β is 20 in our simulations, and much of the analysis is based on Nash equilibrium considerations that implicitly assume infinitely strong selection. But behavior that deviates from the Nash equilibrium can predominate (even without extra mechanisms such as information) when selection is weak [18, 21, 22]. A study examining the trust game in the case of weak selection—and in particular examining whether weak selection is a plausible explanation for human behavior in anonymous one-shot trust games—would be useful in determining the ultimate reasons for human trust.

Even if trust arose in a way unrelated to information, however, our results have shown the powerful effect that information does have when it is available. Given that information can be introduced exogenously—by governments, marketplaces, and the like—our work has important implications for the promotion of trust and, therefore, economic well-being [8, 9, 13, 15, 24], wherever it may be lacking.

Acknowledgements Both authors were supported by Foundational Questions in Evolutionary Biology Prize Fellowships from the John Templeton foundation.

References

- 1. Arrow KJ (1974) The limits of organization. Norton, New York
- Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. Games Econ Behav 10:122– 142
- 3. Bohnet I, Huck S (2004) Reputation and repetition: implications for trust and trustworthiness when institutions change. Am Econ Rev 94:362–366
- 4. Bohnet I, Zeckhauser RJ (2004) Trust, risk and betrayal. J Econ Behav Organ 55:467–484
- 5. Cox JC (2004) How to identify trust and reciprocity. Games Econ Behav 46:260-281
- Fehr E, Kirchsteiger G, Riedl A (1993) Does fairness prevent market clearing? An experimental investigation. Q J Econ 108:437–460
- 7. Glaeser EL, Laibson DI, Scheinkman JA, Soutter CL (2000) Measuring trust. Q J Econ 115:811-846
- Guiso L, Sapienza P, Zingales L (2004) The role of social capital in financial development. Am Econ Rev 94:526–556
- Guiso L, Sapienza P, Zingales L (2006) Does culture affect economic outcomes? J Econ Perspect 20:23– 48
- 10. Guth W, Kliemt H (2000) Evolutionarily stable cooperative commitments. Theory Decis 49:197-221
- 11. Johnson ND, Mislin AA (2011) Trust games: a meta-analysis. J Econ Psychol 32:865–889
- 12. King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR (2005) Getting to know you: reputation and trust in a two-person economic exchange. Science 308:78–83
- Knack S, Keefer K (1997) Does social capital have an economic payoff? A cross-country investigation. Q J Econ 112:1251–1288
- 14. Kreps DM (1990) Corporate culture and economic theory. In: Alt JE, Shepsle KA (eds) Perspectives on positive political economy. Cambridge University Press, Cambridge, p 280
- La Porta R, Lopez-de Silanes F, Shleifer A, Vishny RW (1997) Trust in large organizations. Am Econ Rev 87:333–338
- Malhotra D (2004) Trust and reciprocity decisions: the differing perspectives of trustors and trusted parties. Organ Behav Hum Decis 94:61–73
- 17. Manapat ML, Nowak MA, Rand DG (2012) Information, irrationality, and the evolution of trust. http://ssrn.com/abstract=2102528
- Manapat ML, Rand DG, Pawlowitsch C, Nowak MA (2012) Stochastic evolutionary dynamics resolve the traveler's dilemma. J Theor Biol 303:119–127

Author's personal copy

- 19. McNamara JM, Stephens PA, Dall SRX, Houston AI (2009) Evolution of trust and trustworthiness: social awareness favours personality differences. Proc R Soc B 276:605–613
- 20. Pfeiffer T, Tran L, Krumme C, Rand DG (2012) The value of reputation. J R Soc Interface. doi:10.1098/rsif.2012.0332
- 21. Rand DG, Nowak MA (2012) Evolutionary dynamics in finite populations can explain the full range of cooperative behaviors observed in the centipede game. J Theor Biol 300:212–221
- 22. Smead R (2008) The evolution of cooperation in the centipede game with finite populations. Philos Sci 75:157–177
- 23. Traulsen A, Pacheco JM, Nowak MA (2007) Pairwise comparison and selection temperature in evolutionary game dynamics. J Theor Biol 246:522–529
- 24. Zak PJ, Knack S (2001) Trust and growth. Econ J 111:295-321