האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

ALTRUISM, PARTNER CHOICE, AND FIXED-COST SIGNALING

by

ANDRIY ZAPECHELNYUK
and RO'I ZULTAN

Discussion Paper # 483 May 2008

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Altruism, Partner Choice, and Fixed-Cost Signalling

Andriy Zapechelnyuk [†] and Ro'i Zultan [‡, §]

First version: May 14, 2008; revised version: July 24, 2008

Abstract

We consider a multitype population model with unobservable types, in which players are engaged in the 'mutual help' game: each player can increase her partner's fitness at a cost to oneself. All individuals prefer free riding to cooperation, but some of them, *helpers*, can establish reciprocal cooperation in a long-term relationship. Such heterogeneity can drive cooperation through a partner selection mechanism under which helpers choose to interact with one another and shun non-helpers. However, in contrast to the existing literature, we assume that each individual is matched with an anonymous partner, and therefore, stable cooperation cannot be achieved by partner selection per se. We suggest that helpers can signal their type to one another in order to establish long-term relationships, and we show that a reliable signal always exists. Moreover, due to the difference in future benefits of a long-term relationship for helpers and non-helpers, the signal need not be a handicap, in the sense that the cost of the signal need not be correlated with type.

Keywords: Mathematical model; handicap principle; heterogeneous population; mutual help

[†] Kyiv School of Economics, 51 Dehtyarivska St, 03113 Kyiv, Ukraine. *Email:* zapechelnyuk@kse.org.ua

[‡] Center for Rationality, Hebrew University, Giv'at Ram, 91904 Jerusalem, Israel. *Email:* roi.zultan@mail.huji.ac.il

[§] Andriy Zapechelnyuk and Ro'i Zultan contributed equally to this paper.

1 Introduction

Altruism, expressed as cooperative behaviour in a prisoner's dilemma type game (such that each player increases her partner's fitness at a cost to oneself), usually involves great benefits for both players. However, altruistic behaviour is not possible to maintain in equilibrium, since each individual would prefer to free ride while her partner does all the work. One solution to the cooperation problem is by repeated interactions (Trivers, 1971; Axelrod and Hamilton, 1981). If the two partners are in a long-term relationship, stable cooperation can be obtained in a subgame perfect equilibrium, as each player would help her partner in anticipation of reciprocity (Aumann and Shapley, 1994; Rubinstein, 1979). However, a crucial condition for cooperation in a long-term relationship is that the partners are patient enough, so that they do not discount future payments above a certain threshold. Otherwise the benefits of today's selfish behaviour might loom larger than the future gains from cooperating (Friedman, 1971; Fudenberg and Maskin, 1986).

The level of patience (or discount rate) may be heterogeneous in the population, due to varying characteristics. For example, current resources may affect a threshold horizon for accumulating additional resources, so that a hungrier individual assigns lower value to a future gain.[1] Imagine, then, a population made of individuals varying in their patience. Such a population can be conceptually divided into potential helpers and non-helpers, such that (only) when two helpers interact, cooperation is sustainable as an equilibrium in a repeated game. In this population, helpers would like to identify each other in order to establish long-term relationships with other helpers while avoiding non-helpers.[2] However, assuming that uncooperative partnerships can be terminated in favour of searching for a more beneficial partnership (at a fixed

[1] Note that, assuming that the heterogeneity has its roots (to some extent) in environmental factors, types are not hereditary. Therefore, similar to Fishman et al. (2001), evolutionary processes will not change the types distribution (cf. Eshel and Cavalli-Sforza, 1982; Batali and Kitcher, 1995; Ohtsuki and Iwasa, 2004, 2006).

[2] In such a situation helping may even evolutionary subside the non-helping strategy to become the prevalent strategy throughout the population (Eshel and Cavalli-Sforza, 1982).

cost), non-helpers would like to mimic helpers, thus luring other helpers into interacting with them. In such a way a non-helper can gain the benefits of the first round exploitation before moving on to find a new 'victim'.

Active partner selection when the (potential) partners can be identified has been suggested as a way to explain various examples of cooperation in natural populations (Dawkins, 1976; Axelrod and Hamilton, 1981; Bull and Rice, 1991). The crucial issue of identifying non-helpers has been conceptually attributed to mechanisms such as parcelling, distributing and image scoring, all of which involve obtaining information about possible partners, either through multiple interactions, or by observing previous actions (Sachs et al., 2004). Various theoretical and numerical models have been studied to show that partner selection can facilitate stable cooperation in a homogeneous population. For example, population simulations show that if an individual can decide whether to continue a long-term relationship with her current partner or break the partnership to look for a new partner, cooperative strategies can prosper, provided that the exogenous probability of continuing the relationship is high enough (Schuessler, 1989). If the population is small enough so that repeated encounters occur sufficiently often, an outside option to social interaction also allows for cooperation to emerge, as defectors would be avoided in the future in favour of the outside option (Batali and Kitcher, 1995). Alternatively, in order to instate stable cooperation, repeated encounters can be replaced with free information flow within the population, so that a defector will be 'ostracized' by the whole population (Hirshleifer and Rasmusen, 1989). [3]

In the framework we wish to study, however, we have different types which are exogenously separated by their preferences. Furthermore, as there is no flow of information, and a once terminated relationship can never be restarted as the population is large enough to neglect the probability of a repeated encounter, each individual observes only her private history. Therefore individuals remain anonymous as regards their type, and thus the various identification mechanisms described above break down.

[3] Such reputation mechanisms can also be the basis for indirect reciprocal strategies, thus facilitating cooperation (e.g., Sugden, 1986; Nowak and Sigmund, 1998; Ohtsuki, 2004; Ohtsuki and Iwasa, 2004, 2006).

Nonetheless, we suggest that the helpers may still be able to identify each other by actively signalling their type in a reliable way, which the non-helpers will not be able to imitate. If such a signal exists, then helpers will establish long-term cooperational relationships with other helpers based on the signal, while the non-helpers are excluded from such cooperation by not sending the signal.

In order for such a signal to be reliable, it must involve a handicap, which traditionally means differential costs to the different types (Zahavi, 1975; Grafen, 1990).[4] However, a costly signal can also be reliable when the different types face differential potential benefits from signalling (cf. Johnstone, 1997). Thus, an individual can use a costly signal in order to honestly relay his need for help to another, genetically related, individual (Maynard Smith, 1991; Johnstone and Grafen, 1991; Reeve, 1997). One example of such costly fixed-cost signalling is offspring's begging for food from feeding parents (Godfray, 1991, 1995). Furthermore, fixed cost signalling in mate-choice situations is possible when there are differential benefits from reproduction (Getty, 1998a,b).

We follow these examples of fixed-cost signalling between relatives or potential mates to show that the repeated interactions between anonymous strangers in our model give rise to fixed-cost signalling. Namely, there exists a costly public signal which carries the same cost for all individuals in the population independent of type, and is instrumental in separating the helpers from the non-helpers. Since it may be hard to find and nigh impossible to create a signal which carries a cost correlated with type, it is easy to assume that a fixed expenditure is easily realized. Thus the model we suggest significantly expands the ways in which stable cooperation in a multitype population may emerge.

The separation stems from the different potential long-term benefits from interaction associated with developing this signal. Consider the case in which the (discounted) benefit that the helpers gain from establishing a long term

[4] The handicap may be itself a display of altruism in a preliminary stage (Roberts, 1998; Gintis et al., 2001) or in separate interactions, such that individuals behave altruistically in one relationship in order to signal their type to a third party (Lotem et al., 2003).

relationship is greater than the (discounted) benefit that a non-helper gains from repeatedly deceiving helpers at the cost of the repeated search for new partners. It is then possible for helpers to develop a signal at a cost which is within this margin, and interact only with other individuals who exhibit the signal, while the non-helpers gain more from not developing the signal, and forgo the potential benefits from interaction.

The paper is structured as follows. Section 2 describes the infinitely repeated mutual helping game. Section 3 introduces the multitype population game, and Section 4 introduces an initial phase of public signal choice and analyzes the resulting signalling game. We generalize the results for a continuum of types in Section 5. Section 6 discusses the importance of the model in relation to the existing literature and concludes.

2 The Mutual Help Game

By a *mutual help game* we understand the following two-player interaction. Each player either 'helps' or 'does not help' the other player. The cost of help is denoted by c, the benefit of receiving help from the other player is denoted by v. We assume $v > c > 0$. The payoff functions of the players are summarized in the following table (Fig. 1).

$1 \setminus 2$	do not help	help
do not help	0, 0	v, $-c$
help	$-c, v$	$v-c, v-c$

Fig. 1. Mutual Trust stage game

This game is the Prisoner's Dilemma, with the social optimum achieved when the players help each other, and with the unique Nash equilibrium when the players do not help each other.

Consider now an infinitely repeated mutual help game. Denote by u_{it} the realized stage payoff of player i at period t. The game payoff of player i is a discounted sum of stage payoffs in all periods, that is, an individual i receives $\sum_{t=1}^{\infty} \delta_i^{t-1} u_{it}$, where $\delta_i \in (0,1)$ is the patience level (discount factor) of player

i.

A long-run mutual help relation (cooperation) is supported in a subgame perfect equilibrium (SPE) if each player i prefers the cooperation to a one-shot advantage over a helping partner, i.e.,

$$(v-c)+\delta_i(v-c)+\delta_i^2(v-c)+\ldots \geq v. \tag{1}$$

To formalize this condition, we shall assume that the cooperation is supported by the *grim trigger strategies*, i.e., each player i helps in the first period and continues to help as long as the partner does the same.[5] Condition (1) is equivalent to $\delta_i \geq c/v$, that is, the cooperation is supported in an SPE if (and only if) both players are patient enough.

3 Population Model with Observable Types

Consider now a large (infinite) population $\mathcal{N}$ of players, consisting of heterogeneous individuals. Every individual i's type is characterized by a patience level δ_i; otherwise all individuals are identical.[6] We assume that there are two types of individuals, H-type and L-type, characterized respectively by δ_h and δ_l, $\delta_h > \delta_l$. The proportion of δ_h-type in the population is $\beta \in [0,1]$.

Every player is matched with another player chosen at random from the population, and the two of them play the mutual help game repeatedly in periods $t = 1, 2, \ldots$, until one of them decides to break the relation. Once the relation is broken, each of these players returns to the 'matching market' and is randomly rematched in the following period. We assume that matching is costly, a player needs to pay a fee to participate in the matching market. However, the

[5] We limit players to helping only under the grim trigger strategy, since in a repeated symmetric Prisoner's Dilemma a cooperative SPE exists if only if cooperation can be supported by grim trigger strategies.

[6] For simplicity of exposition we assume that the types are differentiated only in the discount factor. The results, however, will remain qualitatively the same if we allow the type differentiation in other parameters (though allowing for multi-dimensional types would present additional complexities, since the existence of a separating equilibrium would require existence of a signal correlated with type).

participation in the game is voluntary, that is, before every round, each player seeking for a match decides whether he wants to participate in the matching market or to stay out and receive zero payoff in the current period.

The types of the players are observable, and the individuals can condition their actions in the mutual help game on the types of their partners.

Formally, in every period $t = 1, 2, \ldots$, the following four-stage game is played.

Entry decision. Each player (available for matching) has a choice to 'enter' (i.e., participate in the *matching market*) or to 'stay out'. A player who stays out obtains the stage payoff zero; a player who enters is called *entrant* and pays a participation fee s, $0 < s < v$.

Matching. All entrants participate in pairwise random matching.[7] Formally, let x_t and y_t be the measures of, respectively, H and L types among entrants, $x_t + y_t \leq |\mathcal{N}|$. Every entrant is matched with an H type with probability $\frac{x_t}{x_t+y_t}$ and with an L type with probability $\frac{y_t}{x_t+y_t}$.

Mutual help game. Every pair of matched players plays the mutual help game and receives the stage payoffs of this game.

Break-up decision. In every matched pair, each player independently decides to 'stay' with the same partner for the next period or to 'break up'. If both players in the pair simultaneously decide to stay, in the next period they immediately play the mutual help game (thus skipping the stages of entry decision and matching and not paying the entry cost). Otherwise, if at least one player breaks up, both player will start with the entry decision in the next period.

We say a player behaves *cooperatively* with respect to type δ if, when matched with a new player, she 'helps' the partner j if and only if the partner's type δ_j satisfies $\delta_j \geq \delta$; if the partner has helped her in return, she continues to help him again in all future periods, as long as the partner does the same; otherwise she 'breaks up' and searches for another partner. A player behaves *non-cooperatively* if, whenever matched with another player, she never helps

[7] Another way to formulate the model without affecting the results would be to assume that matching is conditional on type, i.e., two individuals can match if and only if each of them agrees to match the given partner's type.

the partner. Note that there always exists an equilibrium where all players behave non-cooperatively. In this paper, however, we are interested in conditions under which cooperative behaviour (with respect to some δ) is sustainable in equilibrium. [8]

We say that a player of type δ is a *helper* if, when matched with another player of the same (or more patient) type, she prefers a long-run mutual help relation to a one-shot advantage over a helping partner; otherwise we say that a player is a *non-helper.* In a long-run mutual help relation, by playing cooperatively a player of type δ receives

$$-s + (v - c) + \delta(v - c) + \delta^2(v - c) + \ldots = -s + \frac{v - c}{1 - \delta},$$

while by a one-period deviation she receives

$$-s + v + \delta\left(-s + \frac{v - c}{1 - \delta}\right).$$

Thus, the cooperative behaviour (with respect to δ) is sustainable if and only if two constraints hold:
(i) the *incentive constraint*

$$-s + \frac{v - c}{1 - \delta} \geq -s + v + \delta\left(-s + \frac{v - c}{1 - \delta}\right), \tag{2}$$

(ii) the *participation constraint*

$$-s + \frac{v - c}{1 - \delta} \geq 0, \tag{3}$$

and each player's patience level is at least δ. Simplifying (2), we obtain

$$\delta \geq c/s. \tag{4}$$

Note that if the incentive constraint holds, inequality (3) also holds: by assumption, $s < v$, hence (4) implies

$$1 - \delta \leq 1 - c/s < 1 - c/v = \frac{v - c}{v},$$

and therefore

$$-s + \frac{v - c}{1 - \delta} > -s + v > 0.$$

[8] I.e., a subgame perfect equilibrium.

Consequently, a player with patience level δ is a helper if and only if (4) holds. The value of the discount factor $\delta^* = c/s$ is a critical level of patience which separates helpers and non-helpers.

We will show that in this game there exists an equilibrium where a helper matched with another helper becomes engaged in the repeated mutual help game forever; a helper matched with a non-helper does not help and breaks the relation immediately after the first interaction. Hence, by entering the matching market a non-helper cannot receive any benefit, and since entering is costly, every non-helper will stay out of the matching market.

Proposition 1. *There exists an equilibrium where a player enters the matching market and plays cooperatively if and only if she is a helper.*

Proof. Consider the following strategies: (i) every helper enters the market and behaves cooperatively with respect to δ^*, i.e., she 'helps' the partner and stays with her for the next period if and only if the partner is a helper; (ii) every non-helper stays out of the matching market, and plays non-cooperatively [9]. This is a subgame perfect equilibrium, since no player is willing to deviate. Indeed, a non-helper cannot benefit by entering the matching market, since he can receive there at most zero, while entrance is costly. The equilibrium payoff of a helper is $-s + \frac{v-c}{1-\delta_h}$, therefore a helper who has entered the matching market cannot benefit by not-helping (i.e., cheating): by the definition of 'helper', $c \leq \delta_h s$, hence
(a) by deviating once she receives

$$v + \delta_h \left(-s + \frac{v-c}{1-\delta_h} \right) \leq v - c + \delta_h \frac{v-c}{1-\delta_h} = \frac{v-c}{1-\delta_h};$$

(b) by deviating once and leaving the matching market she receives v, and $v \leq \frac{v-c}{1-\delta_h}$ if and only if $\delta_h \leq c/v < c/s$ (the latter inequality is by assumption $s < v$). Also, a helper cannot benefit by staying out of the market, since $-s + \frac{v-c}{1-\delta_h} \geq 0$ if and only if $v - c \geq s - \delta_h s$. This inequality holds, since by assumption, $v > s$, and by the definition of 'helper', $c \leq \delta_h s$. **End of Proof**.

The most interesting situation, which we will focus upon, is when the population contains both helpers and non-helpers, $\delta_l < c/s \leq \delta_h$. Then, when types

[9] That is, the non-helper's off-equilibrium play, if he enters the matching market, is non-cooperative.

are observable, there always exists an equilibrium where every helper becomes engaged in a long-term cooperation with another helper, and every non-helper stays out. Thus, partner selection based on the (observable) type enables the helpers to separate themselves from the non-helpers and to establish long-term cooperative relationships.

4 Hidden Types And Signalling

Let us now assume that players cannot observe each other's type. Every player observes only past actions of herself and her partners. Thus, if two players have repeatedly interacted for a few consecutive periods, one may infer the partner's type by observing the past actions in their repeated interaction. However, when the relation is broken, each player is matched again with a completely unknown [10] partner and has to start from scratch.

In this scenario the partner selection mechanism of the previous section breaks down, as there is no basis for a priori discrimination. Therefore the direct knowledge of types is replaced by active signalling, in the following way. Before period 1, each player is allowed to send a public signal in order to communicate her type. We assume that the signal costs $F \geq 0$ to a sender and (presumably) communicates that the sender is a 'helper'. The signal may be sent only once (it is public and irreversible), it is observed by everyone, and the individuals can condition their actions in the mutual help game on the signals of their partners (in contrast to the previous section, where the actions were conditional on the partner's type).

Suppose that the population contains both helpers and non-helper, $\delta_l < c/s \leq \delta_h$. An equilibrium is said to be *separating* if there exists a signal cost F such that every helper sends the signal and plays cooperatively, while every non-helper does not send the signal and stays out of the matching market.

Note that the separation is on both signalling behaviour and further play. A separation on signalling only is not sufficient to induce different play: there is a trivial equilibrium in which every helper sends a costless signal, every non-

[10] In the infinite population, the probability be randomly matched with the same individual more than once is zero.

helper does not (thus the signal is effective in communication of the types), but in the subgame after the signalling stage, everyone behaves non-cooperatively.

Next, note that existence of a separating equilibrium requires costly communication of the type. Indeed, if the cost of the signal is small, then a non-helper can mimic the helper's signal and enter the matching market. Since he will be matched with a helper with probability 1 (as only helpers are supposed to enter the market in a separating equilibrium), and since by assumption $s < v$, his stage payoff is $-F - s + v > 0$ for small enough F.

Theorem 1. *A separating equilibrium always exists. The cost F of the signal in every separating equilibrium satisfies*

$$\frac{v-s}{1-\delta_n} \leq F \leq -s + \frac{v-c}{1-\delta_h}. \tag{5}$$

Before proving the theorem, let us examine condition (5) on the cost of the signal. Since a helper's payoff in a separating equilibrium is $-F + \frac{v-c}{1-\delta_h}$, the right-hand side inequality in (5) is a helper's participation constraint, it requires that a helper receives a nonnegative payoff. Since by sending the signal and entering the matching market a non-helper receives

$$-F + (-s+v) + \delta_n(-s+v) + \ldots = -F + \frac{v-s}{1-\delta_n},$$

the left-hand side inequality in (5) is a non-helper's incentive compatibility constraint: it requires that a non-helper has no incentive to mimic a helper.

Proof of Theorem 1. First, we show that the interval $\left[\frac{v-s}{1-\delta_n}, -s + \frac{v-c}{1-\delta_h}\right]$ is nonempty, and thus there exists F such that condition (5) holds. Using $\delta_h \geq c/s > \delta_n$, we obtain

$$-s + \frac{v-c}{1-\delta_h} \geq -s + \frac{v-\delta_h s}{1-\delta_h} = \frac{v-s}{1-\delta_h} > \frac{v-s}{1-\delta_n}.$$

Next we show that if F satisfies (5), no player is willing to deviate. As noted above, the left-hand side inequality in (5) requires that a non-helper has no incentive to mimic a helper; the right-hand side inequality in (5) stipulates that a helper does not benefit by staying out of the matching market. Also, by the definition of 'helper', a helper cannot benefit by non-helping (cheating). It is straightforward to verify that any other deviation is either equivalent or

inferior to the above three possibilities. **End of Proof**.

5 A Continuum of Types

The model easily extends to the situation where the set of players' types (levels of patience) is a continuum.

Suppose a player has a random type in $(0,1)$. Let $H : [0,1] \to [0,1]$ be a continuous distribution function, associating with every $x \in [0,1]$ the measure of players in the population with patience level less than x, or, in other words, the probability that a player chosen at random has type $\delta < x$,

$$H(x) = \Pr[\delta < x].$$

The players in the population are engaged in the signalling game as in the previous section.

Let $\delta^* = c/s$. As before, every player with type at least δ^* is a helper, otherwise she is a non-helper. Assume that $0 < H(\delta^*) < 1$, that is, positive measures of both helpers and non-helpers are present in the population.

The following theorem states that a separating equilibrium exists and is supported by a unique separating signal.

Theorem 2. *There exists a unique separating equilibrium, where the cost F of the separating signal given by*

$$F = \frac{v-s}{1-\delta^*}. \tag{6}$$

Proof. Since F satisfies (5), existence is easily verified as in the proof of Theorem 1. To prove uniqueness, let us show that with cost of signal $F' \neq F$ either a helper or a non-helper is willing to deviate.

Assume there is a separating equilibrium with $F' > F$. Let δ' be such that $F' = \frac{v-s}{1-\delta'}$. Solving for δ', we obtain $\delta' > \delta^*$, since

$$\delta' = 1 - \frac{v-s}{F'} > 1 - \frac{v-s}{F} = 1 - (1-\delta^*) = \delta^*.$$

By assumption, H is continuous, hence there is a positive measure of helpers with types in $[\delta^*, \delta')$. Consider a helper with type $\delta \in [\delta^*, \delta')$. By the definition

of 'helper', $-s + \frac{v-c}{1-\delta} \leq \frac{v-s}{1-\delta}$, and since $\delta' > \delta$, we have

$$-s + \frac{v-c}{1-\delta} < \frac{v-s}{1-\delta'} = F'.$$

By playing cooperatively, this player receives $-F' - s + \frac{v-c}{1-\delta} < 0$, thus she would be better off by sending no signal and staying out of the matching market.

Next, assume there is a separating equilibrium with $F' < F$. Again, let δ' be such that $F' = \frac{v-s}{1-\delta'}$, and thus $\delta' < \delta^*$. By assumption, H is continuous, hence there is a positive measure of non-helpers with types in (δ', δ^*). A non-helper with type $\delta \in (\delta', \delta^*)$ does not send the signal and receives the payoff of zero, but, since $\delta' < \delta$, by mimicking a helper he would better off by receiving

$$-F' + \frac{v-s}{1-\delta} = -\frac{v-s}{1-\delta'} + \frac{v-s}{1-\delta} > 0.$$

End of Proof.

6 Discussion

In this paper we analyze a heterogeneous population model combining partner choice and signalling. This model is innovative in linking partner choice and signalling mechanisms, by showing, on the one hand, how signalling can support partner choice, and, on the other hand, how partner choice supports the existence of fixed-cost reliable signalling, a phenomenon which so far received little attention in the literature (cf. Getty, 1998a,b). [11] Unlike previous partner choice models discussed in the literature, in our model players cannot guarantee stable cooperation, as we assume an exogenously determined distribution of types and complete anonymity of new partners. We show that when the population is heterogeneous and large enough (to ensure anonymity), the partner selection mechanism is insufficient to allow potentially cooperative types to achieve cooperation. These assumptions are highly plausible, as environmental and personal factors, such as a local shortage in resources, hunger

[11] Previous studies have combined partner choice and signalling in a different way, such that the behaviour resulting from the partner choice dynamic serves a signal for future interactions (Lotem et al., 2003). In our model, however, initial partner choice is made on the basis of preceding signals.

or distress, may make some individuals rather 'impatient' and, therefore, unwilling to cooperate (see also Fishman et al., 2001).

To solve this problem created by anonymity of types, we suggest that the individuals who are planning to cooperate will try to signal to each other their intentions. We show that reliable signalling emerges endogenously from the model; moreover, due to the difference in expected benefits from cooperation, the cost of this signal need not be correlated with the type. Thus, our model illustrates the possibility of fixed-cost signalling, which is somewhat underrepresented in the biological literature. Getty (1998a) states that "It is widely accepted that a requirement for honest handicap signalling is that higher-quality signallers pay lower marginal costs for advertising" and that "The inference that differential costs explain handicap signalling pervades theoretical discussions and guides experimental tests," and proceeds to demonstrate fixed-cost reliable signalling in a sexual choice context (Getty, 1998b, p. 127, 128). We expand on the previous work by showing how fixed-cost signalling can be reliable in mutual help interactions between unrelated individuals, and establishing the importance of the long-term dynamics and their relationship to the signalling problem. The implications for the theoretical and empirical study of altruism are significant, since a fixed-cost signal is easy to produce, as such a signal can be realized by any wasteful behaviour. In this sense it is equivalent to 'burned money', as discussed in the economic literature (Austen-Smith and Banks, 2000). It is worth noting that a more specific, and somewhat similar fixed-cost signal supported by a repeated interactions mechanism was introduced in the economic literature in the context of costly advertisement for experience. In this case, the producer of a high quality product expects repeated sales to returning satisfied customers, and can signal this by 'burning money' in an expensive advertising campaign, which the producer of a low quality product cannot afford (Nelson, 1970; Milgrom and Roberts, 1986).

The results of our model also illustrate the importance of certain parameters in the model. Most significantly, we see the positive effect of the *participation fee* s on cooperation. Although it intuitively appears that difficulties in finding new partners may hinder the formation of profitable relationships, our model demonstrates reverse effects. Firstly, a greater participation fee makes repeated exploitation more costly, leading to a lower threshold for being a helper. In

essence, more impatient individuals become helpers by necessity. In addition, when we introduce type anonymity and signalling, a greater participation fee reduces the signal cost through a two-fold effect: directly, as the future benefits from pretending to be a helper in order to exploit helpers go down with the related costs, and indirectly, through changing the threshold for helpers. The change of the threshold affects the signal cost, since, as a participation fee goes up and more non-helpers become helpers, the remaining non-helpers are by definition more impatient, and thus stand to gain less from future interactions. [12]

Acknowledgements

This paper was conceived during the activities of the Game Theory and Evolution group in 2006-2007, organized by Sergiu Hart and Avi Shmida and hosted by the Institute for Advanced Studies and the Center for Rationality, the Hebrew University. The authors are thankful to the group members, in particular to its organizers and the host institutions.

References

Aumann, R. J. and L. Shapley (1994). Long-term competition: a game-theoretic analysis. In N. Megiddo (Ed.), *Essays in Game Theory in Honor of Michael Maschler*, pp. 1–15. Springer, New York.

Austen-Smith, D. and J. S. Banks (2000). Cheap talk and burned money. *Journal of Economic Theory 91*, 1–16. DOI: 10.1006/jeth.1999.2591.

Axelrod, R. and W. D. Hamilton (1981). The evolution of cooperation. *Science 211*, 1390–1396. DOI: 10.1126/science.7466396.

Batali, J. and P. Kitcher (1995). Evolution of altruism in optional and compulsory games. *Journal of Theoretical Biology 175*, 161–171. DOI: 10.1006/jtbi.1995.0128.

Bull, J. J. and W. R. Rice (1991). Distinguishing mechanisms for the evo-

[12] This two-fold effect is evident in the left hand side of (5), as s appears in the numerator, and δ_n in the denominator.

lution of cooperation. *Journal of Theoretical Biology 149*, 63–74. DOI: 10.1016/S0022-5193(05)80072-4.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Eshel, I. and L. L. Cavalli-Sforza (1982). Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences of the USA 79*, 1331–1335. DOI: 10.1073/pnas.79.4.1331.

Fishman, M. A., A. Lotem, and L. Stone (2001). Heterogeneity stabilizes reciprocal altruism interactions. *Journal of Theoretical Biology 209*, 87–95. DOI: 10.1006/jtbi.2000.2248.

Friedman, J. W. (1971). A noncooperative equilibrium for supergames. *Review of Economic Studies 38*, 1–12. DOI: 10.2307/2296617.

Fudenberg, D. and E. Maskin (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica 54*, 533–554. DOI: 10.2307/1911307.

Getty, T. (1998a). Handicap signalling: when fecundity and viability do not add up. *Animal Behaviour 56*, 127–130. DOI: 10.1006/anbe.1998.0744.

Getty, T. (1998b). Reliable signalling need not be a handicap. *Animal Behaviour 56*, 253–255. DOI: 10.1006/anbe.1998.0748.

Gintis, H., E. Smith, and S. Bowles (2001). Costly signaling and cooperation. *Journal of Theoretical Biology 213*, 103–119. DOI: 10.1006/jtbi.2001.2406.

Godfray, H. C. J. (1991). Signalling of need by offspring to their parents. *Nature 352*, 328–330. DOI: 10.1038/352328a0.

Godfray, H. C. J. (1995). Signalling of need between parents and offspring: parent-offspring conflict and sibling rivalry. *American Naturalist 146*, 1–24. DOI: 10.1086/285784.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology 144*, 517–546. DOI: 10.1016/S0022-5193(05)80088-8.

Hirshleifer, D. and E. Rasmusen (1989). Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior and Organization 12*, 87–106. DOI: 10.1016/0167-2681(89)90078-4.

Johnstone, R. A. (1997). The evolution of animal signals. In J. R. Krebs and N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach*, pp. 155–178. Blackwell, Oxford.

Johnstone, R. A. and A. Grafen (1991). The continuous Sir Philip Sidney

game: a simple model of biological signalling. *Journal of Theoretical Biology 156*, 215–234. DOI: 10.1016/S0022-5193(05)80674-5.

Lotem, A., M. A. Fishman, and L. Stone (2003). From reciprocity to unconditional altruism through signalling benefits. *Proceedings of the Royal Society of London, Series B: Biological Sciences 270*, 199–205. DOI: 10.1098/rspb.2002.2225.

Maynard Smith, J. (1991). Honest signalling: the Philip Sidney game. *Animal Behaviour 42*, 1034–1035. DOI: 10.1016/S0003-3472(05)80161-7.

Milgrom, P. R. and J. Roberts (1986). Price and advertising signals of product quality. *Journal of Political Economy 94*, 796–821. DOI: 10.1086/261408.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy 78*, 311–329. DOI: 10.1086/259630.

Nowak, M. A. and K. Sigmund (1998). Evolution of indirect reciprocity by image scoring. *Nature 393*, 573–577. DOI: 10.1038/31225.

Ohtsuki, H. (2004). Reactive strategies in indirect reciprocity. *Journal of Theoretical Biology 227*, 299–314. DOI: 10.16/j.jtbi.2003.11.008.

Ohtsuki, H. and Y. Iwasa (2004). How should we define goodness? – reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology 231*, 107–120. DOI: 10.1016/j.jtbi.2004.06.005.

Ohtsuki, H. and Y. Iwasa (2006). The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology 239*, 435–444. DOI: 10.1016/j.jtbi.2005.08.008.

Reeve, H. K. (1997). Evolutionarily stable communication between kin: a general model. *Proceedings of the Royal Society of London, Series B: Biological Sciences 264*, 1037–1040. DOI: 10.1098/rspb.1997.0143.

Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society of London, Series B: Biological Sciences 265*, 427–431. DOI: 10.1098/rspb.1998.0312.

Rubinstein, A. (1979). Equilibrium in supergames with the overtaking criterion. *Journal of Economic Theory 21*, 1–9. DOI: 10.1016/0022-0531(79)90002-4.

Sachs, J. L., U. G. Mueller, T. P. Wilcox, and J. J. Bull (2004). The evolution of cooperation. *Quarterly Review of Biology 79*, 135–160. DOI: 10.1086/383541.

Schuessler, R. (1989). Exit threats and cooperation under anonymity. *Journal of Conflict Resolution 33*, 728–749. DOI: 10.1177/0022002789033004007.

Sugden, R. (1986). *The Economics of Rights, Co-Operation, and Welfare*. Blackwell, Oxford.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology 46*, 35–57. DOI: 10.1086/406755.

Zahavi, A. (1975). Mate selection – a selection for a handicap. *Journal of Theoretical Biology 53*, 205–214. DOI: 10.1016/0022-5193(75)90111-3.