Journal of Economic Behavior and Organization 203 (2022) 368–381

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Strategic reciprocity and preference formation

Jose A. Carrasco[a,*], Rodrigo Harrison[a], Mauricio G. Villena[b]

[a] Escuela de Negocios, Universidad Adolfo Ibáñez, 2640 Diagonal Las Torres, Santiago, Chile
[b] Facultad de Administración y Economía, Universidad Diego Pórtales, Avenida Santa Clara 797, Huechuraba, Santiago, Chile

ARTICLE INFO

Article history:
Received 18 July 2021
Revised 5 July 2022
Accepted 11 September 2022

JEL classification:
C72
A13

Keywords:
Reciprocity
Preference formation
Altruism

ABSTRACT

We model how individual preferences are shaped by strategic reciprocity choices. Our model accounts for heterogeneous players — with intrinsic altruistic, selfish or spiteful preferences —who randomly engage in short-run, as well as long-run, pairwise interactions. To disentangle the strategic component of preferences we allow players to act reciprocally in the long-run to conveniently adjust their preferences depending on who they interact with. How they change and what specific kind of preferences emerge in equilibrium crucially depend on whether the short-run strategic interaction is one of strategic complements or substitutes. Our model also predicts that we might observe *behavior-reversion*: players might behave against their intrinsic type exclusively due to strategic considerations. With incomplete information the strategic component of preferences vanishes, equilibrium preferences are as selfish as possible, and thus there is no behavior-reversion.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

People adjust their preferences and behavior after they see what others do. In human relations, as individuals constantly interact with others, many times this behavioral response is simply based on biased perceptions of intentions or how intentions are judged. As such, it is common to observe an individual behaving altruistically in some cases, while selfishly in others. Indeed, although cooperation and altruism lie at the heart of human lives, there are many other situations in which we lash out and choose to do harm. Recognizing these regularities in human interactions, the economic literature has extensively studied preference formation and altruism, but little is known about the internal drivers that shape these altruistic preferences. In this paper we aim to understand how endogenous and *strategic reciprocity* choices shape preferences. Specifically, how a person's desire to reciprocate kindness and spitefulness with in-kind actions influences the concern they show for the well-being of others.

We propose a model where players strategically adjust their preferences and behavior influenced by who they interact with. More specifically, we account for pairwise random meetings between heterogeneous players that inherit a fixed *intrinsic preference type* — altruistic, selfish, or spiteful — and that engage in a simultaneous-move *short-run extraction game* (Okuno-Fujiwara and Postlewaite, 1995). In each meeting, players' intrinsic types summarize underlying intentions, and as such might not necessarily be common knowledge. Of course, players have to be aware about their own intentions, but not necessarily about others. Crucially, and despite the fact that preferences are fixed in the short run, they can be modified in the *long-run* when players engage in strategic interaction at the preference level. The idea of individuals choosing

* Corresponding author.
E-mail addresses: jose.carrasco@uai.cl (J.A. Carrasco), rodrigo.harrison@uai.cl (R. Harrison), mauricio.villena@udp.cl (M.G. Villena).
URL: http://www.tonocarrasco.com (J.A. Carrasco), http://www.rodrigoharrison.com (R. Harrison), http://www.mauriciovillena.com (M.G. Villena)

https://doi.org/10.1016/j.jebo.2022.09.008
0167-2681/© 2022 Elsevier B.V. All rights reserved.

preferences can be thought as a dual-self problem, as proposed by Coleman (1990). Long-run preferences are chosen by an "inner" self, and later acted upon by an "outer" self in the short-run that takes them as given. We make this distinction between short-run and long-run to account for a person's ability to influence or restrict their future decisions, which could even mean pursuing other objectives. Furthermore, our distinction between intrinsic and induced preferences is inspired by Levine (1998) and tries to capture reciprocity in the sense that people are willing to be more altruistic to an opponent who is more altruistic toward them. Our key model ingredient is that the way in which preferences change is through reciprocity. That is, in each meeting players decide the weight on each other's intrinsic type or underlying intentions. Crucially then, preferences are neither given, nor evolutionarily selected before they become givens. Instead, the desire to reciprocate is an endogenous long-run strategic consideration and thus, the main driver that shape preferences. To account for different strategic contexts, we introduce a model parameter $-1 \leq k \leq 1$, where $|k|$ captures the *degree of strategic interaction* in the short-run game. We refer to the environment as one of negative externalities and strategic substitutes when $k < 0$, and one of positive externalities and strategic complements when $k > 0$.[1] As we highlight throughout the paper, the type of game that players engage in (whether is one of complements or substitutes) plays a major role in our results. For it also has long-run consequences and modifies the way in which players reciprocity choices relate. In particular, when the short-run game is one of strategic complements (substitutes), long-run reciprocity choices endogenously become strategic substitutes (complements). Furthermore, the fact that the type of strategic interaction endogenously reverses in the long-run fully determines the kind of preferences that arise in equilibrium. Specifically, in games of complements reciprocity choices translate into reinforcing mutual concern which in equilibrium favors altruistic behavior. Instead, in games of substitutes reciprocity translates into non-reinforcing mutual concern favoring lower concern and spite.

Allowing for endogenous reciprocity captures at least two key economic insights and offers several testable predictions. First, that people's desire to repay kindness with kindness can be a purely cynical, strategic choice. That is, despite the fact that intrinsic preference types are fixed, players do decide to adjust their preferences exclusively due to strategic motives. Indeed, we formally show how players optimally decide to reciprocate and thus how they optimally adjust their preferences due to strategic considerations. We deduce that in general players act differently than what their intrinsic preference type would have suggested, and characterize reciprocity choices as well as the kind of preference that are induced in equilibrium (Proposition 1).

Naturally, if both players' intrinsic preference types coincide (both are altruist or both are spiteful), despite the strategic component that influence their choices, players can not go against their nature. Intrinsically altruistic players behave altruistically, while spiteful ones act spitefully, regardless of the type of game that is played. However, as we account for intrinsic types we are able to show that spite and altruism emerge at its lowest intensity in games of complements and substitutes, respectively. This insight is in stark contrast to previous theoretical work on endogenous preferences, which mainly predict that positive concern for others (altruism) arises only when the strategic context is one of the strategic complements; otherwise, only negative concern (spitefulness) will arise (Bester and Güth, 1998; Bolle, 2000; Possajennikov, 2000; Carrasco et al., 2018). Furthermore, when the intrinsic types of players differ significantly among themselves (i.e., only one is an altruist or only one is spiteful), then altruism can only emerge in games of strategic complements; otherwise, in games of strategic substitutes only spiteful preferences can emerge. That is, in this case at least one player will necessarily behave against his intrinsic type.

For a concrete example as to how strategic reciprocity choices offers new economic insights, consider the *public good contribution game* presented in Levine (1998). In the two-player version of this game, players make a simultaneous and independent costly donation to a common pool. Due to free riding, with selfish players it is a dominant strategy not to contribute at all, but as the authors find in experiments, when players are sufficiently altruistic they do decide to contribute. However, this is mainly because in their model both reciprocity and the intrinsic preference of each opponent are fixed parameters. Instead, in our model we allow players to choose how much to reciprocate depending on with whom they interact; this is a key ingredient that constitutes a radical difference with respect to Levine (1998). As we formally show in our Appendix A.2 for this particular example, when we account for our model ingredients then no player will choose to behave sufficiently altruistically. Hence, no one cooperates; unless, of course, the intrinsic types restrict this choice. These are new and different predictions with respect to the observables (i.e., how much to cooperate) that only arise as a consequence of our strategic reciprocity choices and endogenous preference formation.

Our second insight is that when these strategic effects are significant enough, in expectation they can offset players' intrinsic types, making them act against their nature.[2] We refer to this phenomenon as *behavior reversion*, and we show that it only occurs for *moderate players*; those whose intrinsic preferences are neither too altruistic nor too spiteful (Proposition 2). The rest of the *extreme players* (i.e., those too altruistic or spiteful), although moderate themselves, do not reverse their behavior. Naturally, the specific types of moderate players that reverse their behavior depends on the strategic context of the short run game; moderate-altruistic players reverse their behavior in games of substitutes and moderate-spiteful in games that exhibit complementarity.

[1] This parameter simultaneously captures several economic applications in the context of extraction games. If $k = -1$ then our model is reduced to a simple version of the well known *common-pool resource game*. If instead $k \neq -1$, then our model represents an extraction game with positive ($k > 0$) or negative ($k < 0$) externalities. In market applications, our parameter k accounts for differentiated goods.

[2] This insight is consistent with experimental work that supports the idea that selfish agents do not necessarily behave in line with pure material self-interest (Güth et al., 1982; Isaac and Walker, 1988; Fehr et al., 2002; Charness and Rabin, 2002; Fehr et al., 1993).

Obviously, the strategic component that drives reciprocity and preference formation depends on the information in the hands of players. We show that this is indeed a crucial ingredient and that when there is incomplete information regarding other player types equilibrium preferences are as selfish as possible, as any strategic consideration vanishes. In fact, the optimal reciprocity choice is a dominant strategy, and players reciprocate by weighting their opponent's expected type (Proposition 3). Interestingly, and unlike the perfect information case, optimal reciprocity is independent of the short-run game strategic environment summarized in the parameter k. In other words, players acts as if there were almost none strategic strategic interaction. To wit, in a context of incomplete information behavior-reversion is not possible (Proposition 4). Intuitively, as now the strategic component no longer drives preference formation, obviously it can no longer reverse player's behavior.

Literature Review: Our framework is related to the literature on interdependent preferences and to the endogenous formation of preferences. Unlike our work, the interdependent preferences approach typically considers exogenously specified contexts or fixed preferences that are not influenced by others' behavior (Sobel, 2005; Fehr and Schmidt, 1999; Güth and Napel, 2006; Charness and Rabin, 2002; Koçkesen et al., 2000; Alger and Weibull, 2013; Isoni and Sugden, 2019). More recently, Carrasco et al. (2018) explore the evolutionary stability of interdependent preferences in a context with perfect information and a strategic environment that shows negative externalities and strategic substitutes. We depart from these previous works because our framework is not evolutionary or cultural transmission based, and it considers players' optimizing behavior and strategic interaction under a rich set of different contexts.

The fundamental experimental work provided by Levine (1998) suggests that individuals behave as if they have reciprocal preferences, a more specific type of interdependent preference. Agents with these preferences adjust the concerns they express for others based on their perceptions of how they are being treated by their opponents. Naturally, behavior might change as players may perceive intentions differently as they interact. We use the linear approach proposed by Levine (1998) to distinguish between intrinsic preferences from equilibrium ones. In an alternative indirect evolutionary approach to explain preferences, Dekel et al. (2007) puts focus instead on the stability and efficiency of outcomes.[3] Unlike them, our focus is the strategic component of reciprocity that shapes preferences. They show that when preferences are observable only efficient outcomes are also stable. Instead, when they are not, all strict equilibria are stable.

There are of course many other different ways to account for interdependent preferences. For instance, and in contrast to our paper, Alger and Weibull (2013) study interdependent preferences when different types of players have a concern for efficiency. Instead, Rabin (1993) develops a model where players have a concern for fairness that is exclusively driven, in contrast to our model, by the underlying belief about intentions derived from observable actions. In Falk and Fischbacher (2006) they consider that people evaluate the kindness of an action not only by its intention but also by its consequences.[4] Ours is an intention-based model that exploit Levine (1998) functional form for preferences and assume that players linearly care about their own as well as their opponents.

Sethi and Somanathan (2001) provide an evolutionary explanation of the emergence and stability of reciprocal preferences. Similar to us, they employ a variation of Levine (1998) preference specification to model preferences. However, they consider only two types of players, spiteful-materialists and altruistic-reciprocators, and provide sufficient conditions for stable preferences. In contrast to them, we model reciprocal preferences more generally as we do not require players to have the same intrinsic preferences. In fact, this heterogeneity is one of our key model ingredients. From a theoretical viewpoint, a recent work by Fershtman and Segal (2018) is another attempt to connect preferences and social influence. Their work also considers a social interaction set up where individual behavior not only depends on one's own preferences but also on the behavior of other agents. They assume the existence of a social influence function that converts the private utility functions of all players into an individual observable utility function. By doing so, they study properties of social influence functions and their equilibrium implications, but without proposing an explicit behavioral model. Unlike us, they do not account for strategic behavior at the preference level. Even if players are aware that they influence others, they do not behave strategically, which is a distinct property of our model.

Finally, there may be similarities in our results with Bester and Güth (1998), as well as with Heifetz et al. (2007a) and Heifetz et al. (2007b). While the results may be reminiscent, our methods, approaches, and assumptions are not identical to those of the aforementioned authors. Bester and Güth (1998) follows an evolutionary game theoretic approach and propose a symmetric model with a measure of altruism more restricted than ours. In Heifetz et al. (2007a) an exogenously given share of matches interact under complete information, while the remaining share interact under incomplete information. Our analysis is of complete and incomplete information separately, meaning we do not have different shares of matches playing different games simultaneously. Additionally, Heifetz et al. (2007b) propose an evolutionary model in which matched individuals might receive noisy signals of the opponents preference; in contrast to them, our focus is the endogenous formation of preferences through strategic choices of reciprocity.

[3] Alternatively, Ok and Vega-Redondo (2001)lay out an evolutionary foundation for individualistic preferences. The literature on preference evolution in social interactions is vast, but Alger and Weibull (2019) provide a comprehensive recent survey on it.

[4] The work of Dufwenberg and Kirchsteiger (2004) and McCabe and Smith (2000), propose additional models in which reciprocity is purely intention-based. Instead, Fehr and Schmidt (1999), Levine (1998) and also Bolton and Ockenfels (2000) are all models where reciprocity is outcome-based.

2. The model

We consider a continuum of players that are randomly matched in pairs. Matched players are indexed with $i, j \in \{1, 2\}$ where $i \neq j$. When matched, each player i independently chooses $x_i \in \mathbb{R}_+$ and derives *material payoffs* $\pi_i(x_i, x_j) = x_i(1 - x_i + kx_j)$; the parameter $|k| \leq 1$ captures the *degree of strategic interaction* between players choices. This payoff function captures the idea that individuals face a social dilemma where the pursuit of individual interest comes at the expense of the collective goals.[5] However, preferences are assumed to be *interdependent* and so each player i chooses x_i to maximize his *perceived utility* $u_i(x_i, x_j) = \pi_i(x_i, x_j) + \beta_{ij}\pi_j(x_j, x_i)$, where β_{ij} summarizes his concern over player j material payoff. We say that player i's preferences are altruistic if $\beta_{ij} > 0$, spiteful if $\beta_{ij} < 0$, and selfish if $\beta_{ij} = 0$. This pairwise interaction defines a *short-run extraction game* whose Nash equilibrium is described by action profile (x_i^*, x_j^*).

As we aim to understand how equilibrium preferences are shaped, our model accounts for two key ingredients. First, for *reciprocity and long-run* strategic interaction at the preference level. More specifically, we endogenize the desire to reciprocate kindness and spitefulness with in-kind actions as a long-run strategic consideration. Second, for player heterogeneity: we let players vary by their *intrinsic preferences type* $\theta_i \in [-1, 1]$ and we refer to them as *intrinsically altruistic, selfish or spiteful* if $\theta_i > 0$, $\theta_i = 0$ or $\theta_i < 0$, respectively. We let $\theta \sim F$ with continuous densities $f \equiv F'$ on $[-1, 1]$ and $\mathbb{E}_\theta[\theta] = \mu$.

We combine the aforementioned ingredients adopting Levine (1998) functional form for preferences letting $\beta_{ij} \equiv \theta_i + \lambda_i(\theta_j - \theta_i)$, where $0 \leq \lambda_i \leq 1$ is a *reciprocity strategy* that puts weight on the known intrinsic types. Equilibrium reciprocity arise as players exclusively pursue their *long-run material payoffs* $\Pi_i = \pi_i(x_i^*, x_j^*)$, which defines a new *long-run game*. We compute the *subgame perfect equilibria* to deduce how much preferences differ from their intrinsic values.

We explore equilibrium reciprocity when types are common knowledge in Section 3, and the alternative case when they are not in Section 4. In the former case we compute the Nash equilibria, while in the latter we compute the *sequential equilibria*.

3. Reciprocity and equilibrium preferences

We first solve for equilibrium reciprocity when matched player types are common knowledge. That is, in each interaction both players not only know their own intrinsic types, but also their opponent's. Following Rotemberg (1994), we proceed in two stages; first solving the short run game in which preferences are fixed and given, and then the long-run game, where reciprocity is optimally chosen. Ultimately, we compute the set of *subgame perfect equilibrium* that is described by (x_i^*, λ_i^*) for each player that guarantees sequential rationality.

SHORT-RUN GAME: In this game, individuals care not only about their own material payoffs but also about the material payoffs of others; their preferences are assumed to be interdependent.[6] In an encounter they choose x_i to maximize utility $u_i(x_i, x_j)$, so we deduce best responses $x_i(x_j) = (1 + kx_j(1 + \beta_{ij}))/2$. Depending on the value of preferences (fixed at this stage) and the value of the degree of strategic interaction, summarized by k, two extreme cases arise. First, if $\beta_{ij} = \beta_{ji} = 1$ and $k = -1$, then $x_i^* + x_j^* = 1/2$, as best responses perfectly overlap. Second, if $\beta_{ij} = \beta_{ji} = k = 1$ then both players' best responses grow linearly without intersecting; there is no equilibrium in this case. Otherwise, if $-1 < k < 1$ or $\min(\beta_{ij}, \beta_{ji}) < 1$, then the unique equilibrium is:

$$x_i^* = \frac{2 + k(1 + \beta_{ij})}{4 - k^2(1 + \beta_{ij})(1 + \beta_{ji})} \tag{1}$$

Easily, this solution obeys $x_i^*, x_j^* \geq 0$ and allows for both negative externalities and strategic substitutes ($k < 0$), or positive externalities and strategic complements ($k > 0$) in the short-run game. This type of game distinction will show to be critical later in our results, and thus, we will highlight it when necessary. We also deduce that x_i^* rises in β_{ji}, so the way in which player j expresses both his altruism and concern towards player i is by allowing him to choose a greater value of x_i^*. However, we also deduce that $x_i^* - x_j^*$ is proportional to $k(\beta_{ij} - \beta_{ji})$, and thus, this is the first time in which the type of strategic interaction is important. Specifically, if the short run game is a game of strategic complements (substitutes), whoever exerts more concern towards his opponent will choose a higher (lower) value of the action x. In equilibrium, material payoffs are:[7]

$$\Pi_i \equiv \pi_i(x_i^*, x_j^*) = \frac{(2 + k(1 + \beta_{ij}))(2 + k(1 - \beta_{ij}(1 + k(1 + \beta_{ji}))))}{(4 - k^2(1 + \beta_{ij})(1 + \beta_{ji}))^2} > 0 \tag{2}$$

[5] This payoff function is sufficiently general to capture this trade-off, as well as allows us to illustrate the main arguments of our analysis. We might think that the decisions x_i and x_j correspond to the amounts extracted by players from a pool of fixed resources whose size is normalized to 1. In a market application, actions x_i and x_j represent quantity choices (Cournot) or prices (Bertrand) and so the value of k determines if products sold are complements ($k > 0$) or substitutes ($k < 0$).

[6] By Levine (1998), the linearity of the utility in the opponents material payoff is a convenient approximation. See also Rotemberg (1994), Bester and Güth (1998), Bolle (2000), Possajennikov (2000), Carrasco et al. (2018) and Dufwenberg and Kirchsteiger (2019).

[7] Since $2 + k(1 + \beta_{ij}) > 0$, to show that $\Pi_i > 0$ we do: as $4 - k^2(1 + \beta_{ij})(1 + \beta_{ji}) > 0$ then $\Pi_i(\beta_{ij}, \beta_{ji}) > 0 \leftrightarrow 2 + k - k\beta_{ij}(1 + k(1 + \beta_{ji})) > 0$. If $0 < k \leq 1$, as $-k(1 + k(1 + \beta_{ji})) < 0$ then $2 + k - k\beta_{ij}(1 + k(1 + \beta_{ji})) > 2 - (1 + \beta_{ji})k^2 > 0$. If $-1 < k < 0$ then $-1 < 1 + k(1 + \beta_{ji}) \leq 1$ and as $\beta_{ij} \in [-1, 1]$ then $-1 \leq -\beta_{ij}(1 + k(1 + \beta_{ji})) \leq 1$ so $-k(1 - \beta_{ij}(1 + k(1 + \beta_{ji}))) \leq -2k < 2$.

Reciprocity (long-run) Game: We now endogenize preferences by allowing players to choose how much to reciprocate. These choices will shape preferences and capture an essential part of human behavior: each player's desire to reciprocate kindness and spitefulness with in-kind actions is a long-run strategic consideration. That is, and in contrast to most models of reciprocal altruism, we do not take this desire as a primitive of the agent's preferences. Instead, we leave this desire as an individual choice.

The idea that players are able to choose reciprocity, and thus preferences, in the long-run, can be thought as a *dual-self problem*. The selfish "inner" self relinquishes control of actions to an "outer" self whose preferences are molded by this inner self. Therefore, an inner self can make the outer self altruistic, spiteful, or selfish, and either of these preferences are taken as given by him in the short-run (Coleman, 1990).[8]

Exploiting (2), if we let players choose their desired altruism without accounting for their intrinsic types, the unique equilibrium profile $\beta_{ij}^{*}=\beta_{ji}^{*}=k/(2-k)$ arises, as in Bester and Güth (1998) evolutionary approach.[9] Crucially, at this point we replace their evolutionary process with a strategic interaction stage in which social influences impose a *match-specific restriction* that arise from $\beta_{ij}\equiv\theta_i+\lambda_i(\theta_j-\theta_i)$. Then, reciprocity choices λ_i and λ_j solve the long run strategic interaction game at the preference level, and intrinsic values impose that $\min(\theta_i,\theta_j)\leq\beta_{ij}\leq\max(\theta_i,\theta_j)$, which in turn allow equilibria beyond the aforementioned Bester and Güth (1998) symmetric case.

Our endogenization of reciprocity captures at least two key insights. First, that the people's desire to repay kindness with kindness can be a purely cynical, strategic choice. In the long run, individuals selfishly pursue material payoffs to evaluate how much to reciprocate by strategically putting weight on each others intrinsic values (that summarize players intentions and kindness), while intrinsic preferences remain unchanged. Second, that this *strategic effect* might be significant enough to offset their natural intrinsic preferences. In other words, those whose intrinsic preference is to be altruistic (spiteful) can behave against their nature due to strategic considerations. We highlight these insights below and throughout the paper.

We now let each player i choose their reciprocity λ_i to maximize material payoffs (2) itself and solve the long-run game. Formally, the optimization problem is:

$$\begin{aligned}&\max_{\lambda_i}\Pi_i\\ \text{s.t.}\quad &0\leq\lambda_i\leq 1\\ &\beta_{ij}=\theta_i+\lambda_i(\theta_j-\theta_i)\end{aligned}$$

Let us highlight that after having chosen how much to reciprocate, players behave in each meeting as though their utility is given by u_i; meaning that both their long-run as well as their short-run preferences are genuine. Let $\kappa\equiv k/(2-k)$, and to avoid trivialities, assume that $\theta_i\neq\theta_j$. Then, when players choose reciprocity their best responses are:

$$\lambda_i(\lambda_j)=\frac{1}{(\theta_j-\theta_i)}\left(\frac{(1+\beta_{ji}(\lambda_j))\kappa(1+2\kappa)}{(1+\kappa)^2+\kappa(1+\beta_{ji}(\lambda_j))}-\theta_i\right)\tag{3}$$

We now deduce how the type of strategic interaction in the short run is also critical in explaining long-run behavior. In particular, when the short-run game is one of strategic complements, the best responses in (3) are decreasing (i.e., $\lambda_i(\lambda_j)$ falls in λ_j), and so reciprocity choices are strategic substitutes, as depicted on the right panel of Fig. 1. The opposite happens when the short-run game is one of strategic substitutes, as shown in Fig. 2. Intuitively, as x_i^{*} rises in β_{ji}, the endogenous long-run preferences β_{ij} and β_{ji} will inherit the strategic complementarity or substitutability from the short-run game, which in turn reverses at the reciprocity level. The latter follows from the fact that whenever β_{ij} rises in the reciprocity choice λ_i then β_{ji} falls in λ_j.[10] As we now argue, this type of strategic interaction is critical and will determine the kind of preferences that arise in equilibrium. We first specifically compute equilibrium reciprocity.[11]

Proposition 1. *If* $(\theta_i-\theta_j)(\theta_j-\kappa)\geq 0$ *then* $\lambda_i^{*}=1$ *and* $\lambda_j^{*}=0$ *is the unique Nash equilibrium; otherwise, equilibrium reciprocity is* $\lambda_i^{*}=(\theta_i-\kappa)/(\theta_i-\theta_j)$ *and* $\lambda_j^{*}=1-\lambda_i^{*}$ *is the unique equilibrium. When* $\kappa=1$, *these equilibria only exist if* $\max(\theta_i,\theta_j)<1$.

Note that in any equilibrium we always have that $\lambda_i^{*}+\lambda_j^{*}=1$, which guarantees that at least one player chooses to reciprocate by putting weight on his opponent's intrinsic preferences type. Obviously, it is not necessarily implied that both

[8] Additionally, while it is not our focus, another interpretation is evolutionary. As stated by Rotemberg (1994), if emotional reactions are guided by genes, natural selection could favor the reproduction of individuals whose emotions change in self-interested ways. Natural selection could favor genes that lead to the imitation of successful behavior. People appear successful when their material payoffs are high, thus preferences of those individuals could be inferred from their behavior.

[9] In Bester and Güth (1998), an affine transformation of β_{ij} is inherited by each agent. An evolutionary process replaces our strategic interaction stage at the long run level. Unlike us, each player's β parameter applies regardless of which agent it is matched with, meaning that concern in Bester and Güth (1998) is not match-specific. In this case player i's best response is $\beta_{ij}=(1+(2+k)k\beta_{ji})/(4+(2-k)k(1+\beta_{ji}))$. Second order conditions for player i optimization is $-k^2(4+(1+\beta_{ji})(2-k)k)^4/4(2+k(1+\beta_{ji}))^2(2-k^2(1+\beta_{ji}))\leq 0$, and so equilibrium $\beta_{ij}^{*}=\beta_{ji}^{*}=k/(2-k)$.

[10] To see this, assume $\theta_j>\theta_i$. Then $\beta_{ij}=\theta_i+\lambda_i(\theta_j-\theta_i)$ rises in λ_i. However, $\beta_{ji}=\theta_j+\lambda_j(\theta_i-\theta_j)=\theta_j-\lambda_j(\theta_j-\theta_i)$ falls in λ_j.

[11] We omit the case $\kappa=-1/3$, as this was explored by Carrasco et al. (2018).

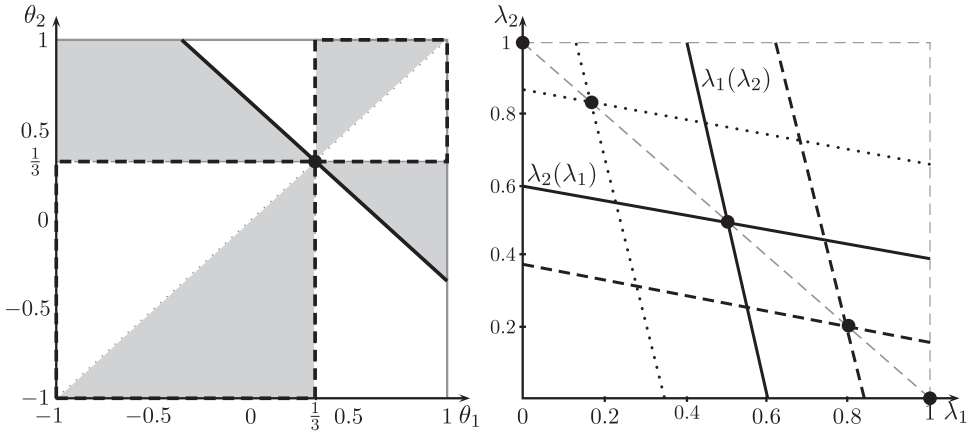

Fig. 1. Preferences and Reciprocity for Complements ($\kappa = 1/3$). Left: Players exert different levels of reciprocity; $\lambda_1^* > \lambda_2^*$ in the gray regions, $\lambda_2^* > \lambda_1^*$ in the white regions and $\lambda_1^* = \lambda_2^*$ along the downward sloping thick line $\theta_1 + \theta_2 = 2/3$. Equilibrium preferences are $\min(\theta_1, \theta_2)$ in the region limited by the upper-dashed square, $\max(\theta_1, \theta_2)$ in the one limited by the lower-dashed one, and equal to κ otherwise. Right: reciprocity choices are substitutes and the best responses slope downward. We use types (θ_1, θ_2) equal to (0.5,0.17) for the solid lines,$(-0.3, 0.5)$ for the dashed lines, and (0.4,0) for the dotted lines. Equilibrium reciprocity choices always obey $\lambda_1^* + \lambda_2^* = 1$.

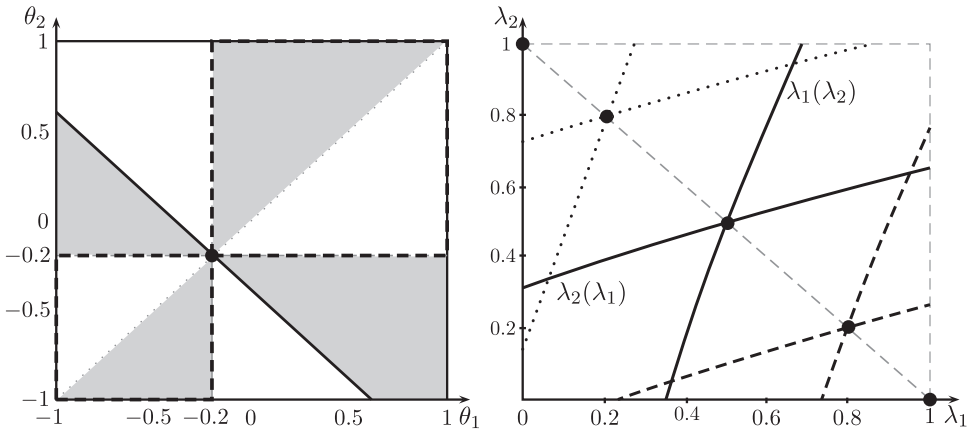

Fig. 2. Preferences and Reciprocity for Substitutes ($\kappa = -1/5$). Left: Players choose $\lambda_1^* > \lambda_2^*$ in the gray regions, $\lambda_2^* > \lambda_1^*$ in the white regions and $\lambda_1^* = \lambda_2^*$ along the downward sloping thick line $\theta_1 + \theta_2 = -2/5$. As in the complements case, equilibrium preferences are $\min(\theta_1, \theta_2)$ in the region limited by the upper-dashed square, $\max(\theta_1, \theta_2)$ in the one limited by the lower-dashed one, and equal to κ otherwise. Right: reciprocity choices are strategic complements and best responses slope upwards. We use types (θ_1, θ_2) equal to $(0, -0.4)$ for the solid lines, $(0, -0.25)$ for the dashed lines, and $(-0.15, -0.4)$ for the dotted lines. Equilibrium reciprocity choices always obey $\lambda_1^* + \lambda_2^* = 1$.

will choose to reciprocate. Regardless, it is always the case that in equilibrium preferences are symmetric.[12] Exploiting the inequality statements of Proposition 1 we now characterize equilibrium preferences:

Corollary 1. *Equilibrium preferences are:*

$$\beta_{ij}^*(\theta_i, \theta_j) = \beta_{ji}^*(\theta_j, \theta_i) = \min(\max(\kappa, \min(\theta_i, \theta_j)), \max(\theta_i, \theta_j))) \tag{4}$$

We now exploit the economics behind our result in Proposition 1, and thus behind the preference specification in (4). We summarize the implications of Corollary 1 in Tables 1 and 2 where we highlight the type of equilibrium preferences that emerge in each encounter and for each type of game.

Intuitively, as preferences are a weighted average of player's intrinsic types, when both individuals are intrinsically altruists then they also have to choose to behave altruistically in equilibrium; they cannot go against their nature. The analog happens if they both were intrinsically spiteful or selfish (see the preferences in the diagonal of Tables 1 and 2). This is in stark contrast to previous theoretical works on preference formation that predict that altruism and spitefulness emerge

[12] In addition, reciprocity is not monotone neither in a players own type nor in the opponents type, by Proposition 1. In particular, it falls in θ_i when $\theta_i \leq \min(\kappa, \theta_j)$ and rises when $\theta_i \geq \max(\kappa, \theta_j)$; otherwise it equals zero, by Proposition 1. Equivalently, it is zero when $\theta_j < \min(\kappa, \theta_i)$ and jumps to one when $\min(\kappa, \theta_i) \leq \theta_j \leq \max(\kappa, \theta_i)$; otherwise falls in θ_j. As for the monotonicity in κ, we see that reciprocity λ_i is a piecewise linear function.

Table 1
Preferences for Complements $\kappa > 0$.

		intrinsic type θ_2		
	Preferences $\beta_{12}^* = \beta_{21}^*$	altruistic $\theta_2 > 0$	selfish $\theta_2 = 0$	spiteful $\theta_2 < 0$
intrinsic type θ_1	altruistic $\theta_1 > 0$	altruistic	altruistic	altruistic
	selfish $\theta_1 = 0$	altruistic	selfish	selfish
	spiteful $\theta_1 < 0$	altruistic	selfish	spiteful

Table 2
Preferences for Substitutes $\kappa < 0$.

		intrinsic type θ_2		
	Preferences $\beta_{12}^* = \beta_{21}^*$	altruistic $\theta_2 > 0$	selfish $\theta_2 = 0$	spiteful $\theta_2 < 0$
intrinsic type θ_1	altruistic $\theta_1 > 0$	altruistic	selfish	spiteful
	selfish $\theta_1 = 0$	selfish	selfish	spiteful
	spiteful $\theta_1 < 0$	spiteful	spiteful	spiteful

in games of strategic complements and substitutes, respectively (Bester and Güth, 1998; Bolle, 2000; Possajennikov, 2000; Carrasco et al., 2018). However, exploiting (4) we also deduce that in games of complements spite emerges at its lowest possible intensity $\max(\theta_i, \theta_j)$. Equivalently, altruism emerges at its lowest possible intensity $\min(\theta_i, \theta_j)$ when the short-run game if one of substitutes.

Otherwise, only when players' intrinsic preferences significantly differ in their type (i.e., only one is an altruist and one is spiteful) the type of game that the individual play in the short-run fully determines preferences. Specifically, as we highlight in Tables 1 and 2, altruistic preferences only arise in equilibrium when the short-run game exhibits strategic complementary of actions (Table 1); otherwise, in games of strategic substitutes (Table 2), equilibrium preferences are spiteful (i.e., $\beta_{ij}^* = \beta_{ji}^* > 0 \leftrightarrow \kappa > 0$).[13] For an intuition of how this result emerges, suppose the short-run game is one of strategic complements ($\kappa > 0$). Furthermore, suppose a meeting between *Ana* ($i = 1$) and *Bob* ($j = 2$), and WLOG suppose $\theta_1 > \theta_2$; that is, Ana is intrinsically more altruistic or less spiteful than Bob. If Bob decides to act more reciprocally by increasing λ_2 then β_{21} will rise as he puts more weight on Ana's type θ_1 that exceeds his. However, λ_1 will fall since reciprocity are strategic substitutes, by (3). To wit, not only will Ana act less reciprocally but β_{12} will rise as she will put more weight on her own type that exceeds Bob's. Ultimately, both Ana and Bob will both increase their concern towards each other. Thus, this mutual concern reinforces, translates into complementarity between the endogenous preferences β's and consequently more altruistic behavior.[14] The type of strategic interaction in the short-run game is also important in explaining how the known and commonly assumed selfish preferences arise in equilibrium (i.e., $\beta_{ij}^* = \beta_{ji}^* = 0$). As preferences in (4) can only take three possible values (κ or either the maximum or the minimum intrinsic type), this will require at least one selfish player, as shown in Tables 1 and 2. It follows then, by simply extending our previous logic, that when intrinsic types differ if the short-run game is one of strategic complements then either altruism or selfishness emerge as equilibrium preferences. However, as we exploit the specifics of (4), selfishness only prevails in meetings between intrinsically selfish and spiteful players. When the short-run game is one of strategic substitutes, only spitefulness or selfishness emerge as equilibrium preferences; in particular, selfishness will only prevail in meetings between intrinsically selfish and altruistic players.

Deeper analysis of our results from Proposition 1, depicted in Figs. 2 and 1, allows us to distinguish the specific reciprocity decisions that induce the preferences in (4). The first thing to note is that symmetric reciprocity choices are unusual; in general, players will choose different reciprocity values. However, this could occur in our model in the particular case where $\kappa = (\theta_i + \theta_j)/2$. In this case we have $\lambda_i^* = \lambda_j^* = 1/2$ and $\beta_{ij}^* = \beta_{ji}^* = (\theta_i + \theta_j)/2$, by Proposition 1. The player that decides to act reciprocally (and also which one more so) crucially depends, again, on the type of short-run game they play. In particular, on how our model parameter κ compares to the average intrinsic type $(\theta_i + \theta_j)/2$. If $\kappa < (\theta_i + \theta_j)/2$ then the most altruistic (or least spiteful) is also more reciprocal (i.e., $\lambda_i^* > \lambda_j^* \leftrightarrow \theta_i > \theta_j$). By the same token, if $\kappa > (\theta_i + \theta_j)/2$ then the less altruistic (or more spiteful) player is more reciprocal (i.e., $\lambda_i^* > \lambda_j^* \leftrightarrow \theta_j > \theta_i$).[15] Re-

[13] Furthermore, the extreme cases of altruism or spitefulness where $\beta_{ij}^* = \beta_{ji}^* = 1$ or $\beta_{ij}^* = \beta_{ji}^* = -1$ are never induced interdependent preferences profiles. As a result, concern for others yield inefficient outcomes. The efficient outcome arises when $\beta_{ij}^* = \beta_{ji}^* = 1$ and so $x_i^* = x_j^* = 1/4(1-k)$ for $k \neq 1$ and $x_i^* + x_j^* = 1/2$ for $k = 1$.

[14] The same logic applies when the short-run game is one of strategic substitutes ($\kappa < 0$), but in this case reciprocity choices are strategic complements. Hence, more reciprocity exerted by Bob yields more reciprocal behavior exerted by Ana. This means that as Bob increases his concern towards Ana (as he puts more weight on Ana's type), she will lower her concern towards Bob (as she puts more weight on Bob's type). Thus, the absence of a reinforcing effect favors lower values of concerned spiteful behavior.

[15] In extreme cases, reciprocity might even be exerted by just one player. This happens, for example, if $\kappa \leq \min(\theta_i, \theta_j)$ or if $\kappa \geq \max(\theta_i, \theta_j)$. In the first case, only the highest type player chooses $\lambda^* = 1$ while the other player is not reciprocal at all and $\lambda^* = 0$; the opposite happens in the second case.

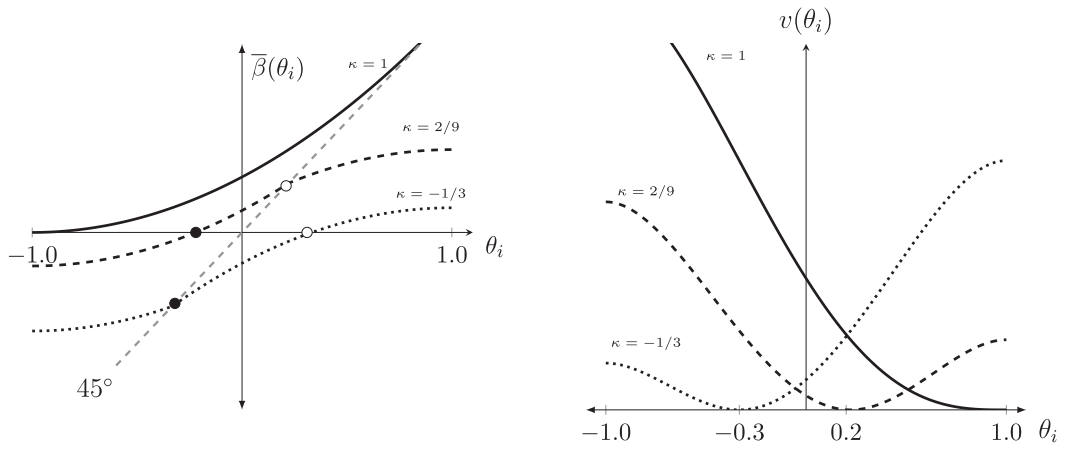

Fig. 3. Expected Preferences and its Variance. Left: we depict the expected preferences $\overline{\beta}(\theta_i)$ and our findings in Proposition 2. For $\theta_i \leq \underline{\theta}$ (black circles) we have $\theta_i \leq \bar{\beta}(\theta_i) \leq 0$, and if $\theta_i \geq \bar{\theta}$ (white circles) then we have $0 \leq \bar{\beta}(\theta_i) \leq \theta_i$. Otherwise, $\bar{\beta}(\theta_i) \leq \min(0, \theta_i)$ if $\kappa < 0$, and $\bar{\beta}(\theta_i) \geq \max(0, \theta_i)$ if $\kappa > 0$. Right: we depict the variance. When $\theta_i = \kappa$, then $\beta^*_{ij} = \kappa$ and thus the variance is zero. We posit uniform distribution for types.

gardless, equilibrium preferences are always symmetric and thus, since $\beta^*_{ij} = \beta^*_{ji} = \beta^*$, we obtain that material payoffs are $\Pi^*_i = (1+\kappa)(1+\kappa(1-2\beta^*))/4(1-\kappa\beta^*)^2$, by (2). That is, altruism increases each player's payoff.[16]

As we previously pointed out, one of the key insights captured by our model is that players can behave against their nature due to strategic considerations; those whose intrinsic preference is to be altruistic (spiteful) could choose to behave spitefully (altruistically). We now explore this idea exploiting our deduced long-run specification for reciprocity and preferences in Proposition 1. While it is true that we could have characterized the difference between equilibrium preferences $\beta^*_{ij}(\theta_i, \theta_j)$ and intrinsic preferences θ_i, this translates into a tedious case-by-case analysis. Instead, and in order to provide clear insights, we compare each player's intrinsic type against their *expected preferences* $\bar{\beta}(\theta_i) = \int \beta^*_{ij}(\theta_i, \omega) dF(\omega)$ that summarize each player i's average concern for their opponent's payoffs. Ultimately, our goal is to understand what behavior is to be expected when players are under social influences and coexist in strategic environments, and how it differs from intrinsic values.[17] Rewriting (4) we have:[18]

$$\bar{\beta}(\theta_i) = \begin{cases} \int \min(\theta_i, \max(\kappa, \omega)) dF(\omega) & \text{for } \theta_i \geq \kappa \\ \int \max(\theta_i, \min(\kappa, \omega)) dF(\omega) & \text{for } \theta_i \leq \kappa \end{cases} \tag{5}$$

Once again, the type of short-run game is critical. Exploiting (5), and as depicted in Fig. 3, we see that $\bar{\beta}(\theta_i)$ rises in θ_i and obeys $\bar{\beta}(\theta_i) \geq \theta_i$ only if $\theta_i \leq \kappa$. That is, not only does the value of κ determine whether altruism or spitefulness emerge in equilibrium, as it can be deduced from Proposition 1. We now see that the comparison between κ and the specific player type exactly determines whether in equilibrium a player expects to behave more altruistically or more spitefully than his true intrinsic value would have dictated. Altogether, when the complementarity in the short run game is sufficiently strong players expect tp behave more altruistically than what they really are; the analogous applies when the substitutability is sufficiently strong.

Our next result formalizes the idea that there might exist cases in which due to large strategic effects players might choose to behave against their nature. That is, that an intrinsically altruistic (spiteful) player expects to behave spitefully (altruistically). We call this phenomenon *behavior-reversion*, whose existence is now established.

Proposition 2. *There is a nonempty set of types where behavior-reversion arises in the long run game.*

This result is based on the identification of two critical values for the intrinsic types that obey $-1 \leq \underline{\theta} < 0 < \bar{\theta} \leq 1$ and define a set $[\underline{\theta}, \bar{\theta}]$ of *moderate intrinsic types*. As depicted on the left panel of Fig. 3, it is precisely for these "moderate players" that behavior reversion might occur. Instead, for *extreme players* with $\theta_i \notin [\underline{\theta}, \bar{\theta}]$ behavior reversals no longer occur.

We highlight three main conclusions derived from Proposition 2, all of them depicted on the left panel of Fig. 3. First, that expected preferences generically differ from intrinsic types; in fact, they only coincide in the particular case when $\theta_i = \kappa$. This suggests that individuals adjust their own preferences due to strategic effects, and thus we should observe them acting differently than what their intrinsic values dictate. Second, due to strategic effects, the extreme players expect to moderate

[16] As $\partial \Pi_i / \partial \beta^* = (1-\beta^*)(1+\kappa)\kappa^2/8(1-\kappa\beta^*)^3 > 0$, this extends Proposition 1 in Bester and Güth (1998). Altruism increases efficiency, whereas spite reduces it.

[17] When $k = -1$, we restrict attention to the symmetric equilibrium, so that in all meetings preferences are symmetric and equal to $\beta^*_{ij}(\theta_i, \theta_j)$, as in (4).

[18] Write $\beta^*_{ij}(\theta_i, \theta_j) = \min(\max(\theta_i, \min(\kappa, \theta_j)), \max(\kappa, \theta_j)))$. As $\theta_i \geq \kappa \geq \min(\kappa, \theta_j)$, the first interval is obvious. Otherwise, since $\theta_i \leq \kappa$ and $\min(\kappa, \theta_j) < \max(\kappa, \theta_j)$, then $\beta^*_{ij}(\theta_i, \theta_j) = \max(\theta_i, \min(\kappa, \theta_j))$.

themselves. In particular, a sufficiently altruistic player $\theta_i \geq \bar{\theta} > 0$ expects to behave altruistically, but not as much as he intrinsically is (i.e., $0 \leq \bar{\beta}(\theta_i) \leq \theta_i$), whereas a sufficiently spiteful player $\theta_i \leq \underline{\theta} < 0$ expects to behave spitefully, but not as much as he is (i.e., $\theta_i \leq \bar{\beta}(\theta_i) \leq 0$). Third, moderate players might reverse their behavior. Furthermore, what specific types of moderate players reverse their behavior depends on the strategic context of the short-run game. In particular, while moderate-altruistic players reverse their behavior when the short-run game is one of strategic substitutes, moderate-spiteful players do so when the game exhibits complementarity.

We also compute the *variance of preferences* $v(\theta_i) = \mathbb{E}_{\theta_j}(\beta_{ij}^{*2}) - \bar{\beta}(\theta_i)^2$ using (4):

$$v(\theta_i) = \begin{cases} 2\int_{\kappa}^{\theta_i}(\theta_i - \omega)F(\omega)d\omega - \left(\int_{\kappa}^{\theta_i} F(\omega)d\omega\right)^2 & \text{for } \theta_i \geq \kappa \\ -2\int_{\theta_i}^{\kappa}(\omega - \kappa)F(\omega)d\omega - \left(\int_{\theta_i}^{\kappa} F(\omega)d\omega\right)^2 & \text{for } \theta_i \leq \kappa \end{cases} \tag{6}$$

As shown on the right panel of Fig. 3 the variance is a U-shaped function of the intrinsic type value; except for when $\kappa = 1$, in which case is decreasing.[19] That is, preferences exhibit more variation when intrinsic types are either more altruistic or more spiteful in comparison with κ. In general, it is hard to provide more insights, as the variance depends not only on a player specific type, but also on the specific shape of the type distribution F.

4. Reciprocity under incomplete information

Unlike our previous section, we now assume that, although players know their own intrinsic type, they do not know their opponent's intrinsic preference. Instead, we assume that players only know the distribution of intrinsic types.

The key economic implication of this new assumption is that now in the long-run game players cannot condition their behavior on who they meet. Instead, each player's reciprocity strategy can only depend on the specific value of their own intrinsic type. Crucially then, as we will formally show, the lack of this valuable information (i.e., their opponents type) immediately eliminates the long-run strategic interaction that guided much of our results in the complete information section.

Regardless, we use the same logic we used in Section 3 to solve for the equilibrium and proceed in two stages.[20]

Short-run game: As we have just stated, players can no longer condition their behavior on who they meet, and their preferences will exclusively depend on their own intrinsic type. Letting $b_i(\theta_i)$ be player i*'s incomplete information preference*, then the *expected utility* is

$$U_i(x_i|\theta_i) = x_i(1 - x_i + k\mathbb{E}_{\theta_j}[x_j(\theta_j)]) + b_i(\theta_i)\mathbb{E}_{\theta_j}[x_j(\theta_j)(1 - x_j(\theta_j) + kx_i)] \tag{7}$$

This is the exact analog of the perceived utility $u_i(x_i, x_j)$ in our previous section except that now the information structure is different. As a result, each player i has to account for his opponent's *Bayesian strategy* $x_j(\theta_j)$, but whose specific intrinsic type is unknown. Solving for the Bayesian equilibrium, we let $\bar{b}_j \equiv \mathbb{E}_{\theta_j}[b_j(\theta_j)]$, and obtain:[21]

$$x_i^*(\theta_i) = \frac{(2 + k(b_i(\theta_i) - \bar{b}_i))(2 + k(1 + \bar{b}_j)) + 2k(\bar{b}_i - \bar{b}_j)}{2(4 - k^2(1 + \bar{b}_i)(1 + \bar{b}_j))} \tag{8}$$

Then, the expected material payoffs $\mathbb{E}_{\theta_j}[\Pi_i] = x_i^*(\theta_i)(1 - x_i^*(\theta_i) + k\mathbb{E}_{\theta_j}[x_j^*(\theta_j)])$ are:

$$\mathbb{E}_{\theta_j}[\Pi_i] = \frac{(4 + k(2 - k\bar{b}_i(1 + \bar{b}_j)))^2 - k^2(2 + k(1 + \bar{b}_j))^2(b_i(\theta_i))^2}{4(4 - k^2(1 + \bar{b}_i)(1 + \bar{b}_j))^2} \tag{9}$$

Our expressions in (8) and (9) are exactly analog to (1) and (2) from our previous section. Comparing them, it is possible to offer some immediate conclusions. First, and most importantly, in the long-run game when players have to choose how much to reciprocate, there will be no strategic interaction. Indeed, inspecting (9) we see that expected material payoffs, to be optimized in the long-run, only depend on each player's preferences $b_i(\theta_i)$, and on other fixed objects such as $\bar{b}_j$ and $\bar{b}_i$. In other words, any desire to reciprocate kindness or spitefulness with in-kind will dissipate. Second, if we do not account for an endogenization of reciprocity then, as expected, we recover the equilibrium found in Bester and Güth (1998), as well as $b_i^*(\theta_i) = \bar{b}_i = k/(2 - k)$. For an intuition, observe that in this case $\bar{b}_i$ and $b_i(\theta_i)$ are both equal and constant terms, independent of the intrinsic type θ_i. Thus (9) is reduced to exactly (2), only when $\beta_{ij} = b_i$ and $\beta_{ji} = b_j$. As a result, the unique equilibrium is obviously $b_i^* = b_j^* = k/(2 - k)$, by Proposition 1. This then highlights the importance to letting reciprocity choices to be endogenous, as in otherwise the equilibrium preferences are the same, regardless of our assumption about the kind of information players have at hand.

[19] For $\theta_i < \kappa$, since $F(\omega) \leq 1$ and thus $\int_{\theta_i}^{\kappa} F(\omega)d\omega \leq \kappa - \theta_i$ we have $\partial v(\theta_i)/\partial\theta_i = 2F(\theta)\left(\theta_i - \kappa + \int_{\theta_i}^{\kappa} F(\omega)d\omega\right) < 0$. For $\kappa < \theta_i < 1$, we have $\partial v(\theta_i)/\partial\theta_i = 2(1 - F(\theta))\int_{\kappa}^{\theta_i} F(\omega)d\omega > 0$. At $\theta = \kappa$ we obtain $v(\kappa) = 0$.

[20] In this case the primary equilibrium concept is sequential equilibrium that imposes besides sequential rationality an additional consistency requirement on beliefs.

[21] Best responses are $x_i(\theta_i) = (1 + k(1 + b_i(\theta_i))\mathbb{E}_{\theta_j}[x_j(\theta_j)])/2$, therefore $\mathbb{E}_{\theta_i}[x_i(\theta_i)] = (1 + k(1 + \bar{b}_i)\mathbb{E}_{\theta_j}[x_j(\theta_j)])/2$. By the same logic, we obtain $\mathbb{E}_{\theta_j}[x_j(\theta_j)]$ and thus $\mathbb{E}_{\theta_j}[x_j(\theta_j)] = (2 + k(1 + \bar{b}_j))/(4 - k^2(1 + \bar{b}_i)(1 + \bar{b}_j))$. We obtain (8) by plugging $\mathbb{E}_{\theta_j}[x_j(\theta_j)]$ in player i's best response.

Reciprocity (long-run) Game: We now solve our long-run reciprocity game when types are not known. Players still act reciprocally, however, unlike the complete information case, they do so by weighting their intrinsic type and the average opponent's type so that preferences are now ($\star$) $b_i(\theta_i) = \theta_i + \lambda_i(\mu - \theta_i)$ with $\min(\theta_i, \mu) \leq b_i(\theta_i) \leq \max(\theta_i, \mu)$.[22]

To solve for the reciprocity strategy we maximize (9) in λ_i, accounting for ($\star$) and taking $\bar{b}_j$ as given. Exploiting our continuum of types assumption, we observe that the specific choice of λ_i does not modify the expected preference $\bar{b}_i$. To wit, we also consider $\bar{b}_i$ as a constant. As a result, we now verify that conditional on their intrinsic preference type, the reciprocity choices are dominant with players reciprocating independent of their opponent's strategy λ_j. To wit, the strategic ingredient of our model vanishes. We formalize this logic in the following Proposition that characterizes the unique sequential equilibrium.

Proposition 3. *For $\theta_i \neq \mu$, the dominant reciprocity strategy for each player i is: (a) $\lambda_i^* = \theta_i/(\theta_i - \mu)$ if $\max(\theta_i, \mu) \geq 0$, (b) $\lambda_i^* = 1$ if $\theta_i > \mu \geq 0$ or $0 \geq \mu > \theta_i$, and (c) $\lambda_i^* = 0$ if $0 \geq \theta_i > \mu$ or $\mu > \theta_i \geq 0$.*

Inspecting (9), as there is only a single term that depends on $b_i(\theta_i)$, we can directly infer that expected material payoffs are maximized when induced preferences $b_i(\theta)$ are as close to as zero as possible. That is, not only is each player's optimal reciprocity strategy dominant, but they also aim for social preferences that are *as selfish as possible*.

Corollary 2. *The expected preferences are:*

$$b_i^*(\theta_i) = \min(\max(0, \min(\theta_i, \mu)), \max(\theta_i, \mu))) \tag{10}$$

Compared to (4), we see that preferences in the incomplete information case are as if players were to match the average opponent's type μ in an environment without strategic interaction (i.e., $\kappa \approx 0$). Thus, players restrict their preferences so that they behave as selfishly as they can, given the constraints imposed by types. Furthermore, not only are reciprocity choices and preferences independent of the opponent's intrinsic type, as one might have expected, but they are also independent of the strategic context summarized in the model parameter κ. This is precisely how our model captures that there is no longer any long-run strategic interaction between players at the preference level.

In terms of observable behavioral predictions, now altruism, spite, and selfishness could all arise in equilibrium. This is in contrast to the complete information case in which the type of short-run game played had a crucial role. Now, the specific type of preferences that emerge crucially depend only on each player's intrinsic type, and the expected type of the opponent. For instance, exploiting (10), we deduce that altruistic preferences are optimal for player i (i.e., $b_i^*(\theta_i) > 0$) if $\min(\theta_i, \mu) > 0$, that spiteful preferences (i.e., $b_i^*(\theta_i) < 0$) are optimal if $\max(\theta_i, \mu) < 0$, and that otherwise selfish preferences (i.e., $b_i^*(\theta_i) = 0$) are optimal if $\mu\theta_i \leq 0$. In words, a player chooses altruistic preferences only if he is altruistic, and if the expected opponent's type he faces also is. By the same logic, if a spiteful player interacts with players that are expected to be spiteful then he will choose spiteful preferences. Selfish preferences only arise for altruistic players that interact with spiteful players, or vice versa. This result is at odds with the findings of Ely and Yilankaya (2001), who for a general model of indirect evolution shows that with incomplete information — when the preferences of the opponent are not known — only egoistic preferences (or preferences equivalent to them) survive evolution.

The divergence of preferences from their intrinsic values is now measured by $b_i^*(\theta_i) - \theta_i$. Exploiting the above inequalities we deduce that, unlike our complete information model, now there is now no behavior-reversion.

Proposition 4. *There is no behavior-reversion in the long run game.*

This result is consistent with the fact that now the strategic component of our model, and thus the effects of social influence on individual's behaviors, is diluted. Consequently, in contrast to our complete information specification, our results show that even under the presence of significant reciprocity exerted by players, they will never behave against their nature. That is, an intrinsically altruistic (spiteful) player will never adjust his preferences towards spiteful (altruistically) behavior.

5. Conclusions

In this paper we aim to understand endogenous reciprocity choices to disentangle how they shape preferences. In order to do so, we formalize the notion that people adjust their preferences and behavior influenced by who they interact with. In our model, players engage in a simultaneous move short-run game and also in long-run strategic interaction at the preference level. Crucially, preferences are neither given, nor evolutionarily selected before they become given. Instead, the desire to reciprocate is an endogenous long-run strategic consideration that determines the formation of preferences.

Our endogenization of reciprocity captures at least two key economic insights and offers several testable predictions. First, as players selfishly pursue material payoffs in the long-run, an individual's desire to reciprocate and repay kindness with kindness can be a purely cynical, strategic choice. We find that the type of short-run strategic interaction (i.e., whether

[22] A natural extension is to account for noisy signals. However, given the linear functional form specified in ($\star$), our results in this section easily extend to the noisy case for in this case $b_i(\theta_i) = \theta_i + \lambda_i(\mu_{ji} - \theta_i)$ where $\mu_{ji} = \mathbb{E}_{\theta_j}(\theta_j|\theta_i)$. To wit, in the special case with additive separable zero-mean noise $\theta_j + \epsilon_i$, we obtain $b_i(\theta_i) = \theta_i + \lambda_i(\mu - \theta_i)$. More sophisticated noise distributions (e.g., type dependent) might be worth to explore; specially with a non-linear preference structure.

is a complements or substitutes game) is critical and has long-run consequences on these players choices. Not only does this determine what kind of preferences emerge in equilibrium, but also how players decide to reciprocate. When intrinsic preferences differ in type, then altruism emerges in games of complements while spitefulness in games of substitutes; selfish preferences might arise in either kind of game. More generally, whenever altruism emerges in games of substitutes it does so at its minimum intensity. The equivalent holds for spite in games of complements.

Second, our model also captures the idea that, due to large strategic effects, people might reverse their behavior and act against their true intrinsic type. In other words, intrinsically altruistic or spiteful players might behave against their nature exclusively due to strategic considerations. We show that only moderate players might reverse their behavior and that extreme players, although they moderate themselves, they do not reverse their behavior. Furthermore, we also show that the strategic context of the short-run game explains the specific types of moderate players that reverse their behavior. While moderate-altruistic players reverse their behavior when the short-run game is one of strategic substitutes, moderate-spiteful players do so when the game exhibits complementarity.

Another takeaway message is that the strategic component that drives reciprocity and preference formation crucially depends on the information in the hands of players. We show that when there is incomplete information regarding the other player's type, the strategic component of reciprocity vanishes. As a result, equilibrium preferences are as selfish as possible and there is no behavior-reversion. Regardless, unlike Ely and Yilankaya (2001), altruistic and spiteful preferences might arise. Consequently, these results seem to suggest that the effects of social influence on individual behavior dilute when there is incomplete information on other player types.

Future extensions to this work include replicating our analytical framework for more general matching technologies, other than pairwise random matching to examine the relationship between group size and reciprocity. Furthermore, our model could be extended to account for heterogeneous flexibility of players in adjusting to strategic concerns. Finally, the theoretical results presented here can give rise to an experimental design to verify how players reverse their behavior or how they reciprocate in context of complete and incomplete information on other player types (for example, showing other players' previous experimental behavior Villena and Zecchetto, 2011).

Appendix A

A1. Omitted Proofs

Proof. Proof of Proposition 1: Consider player $\theta_i \neq \theta_j$ maximization. As the Kuhn-Tucker FOC are necessary, we set up a Lagrangian $\mathcal{L} = \pi_i(x_i^*, x_j^*) + \gamma_0\lambda_i + \gamma_1(1-\lambda_i)$, where $\gamma_0, \gamma_1 \geq 0$ are the multipliers for $\lambda_i \geq 0$ and $\lambda_i \leq 1$. The FOC are:

$$\frac{(2+k(1+\beta_{ji}))((1+\beta_{ji})(2+k)k-\beta_{ij}(4+k(1+\beta_{ji})(2-k)))}{(4-k^2(1+\beta_{ij})(1+\beta_{ji}))^3} = \frac{(\gamma_1-\gamma_0)}{k^2(\theta_j-\theta_i)} \tag{A.1}$$

with $\gamma_0\lambda_i = 0$ and $\gamma_1(1-\lambda_i) = 0$. When $\lambda_i^* = 1$ and $\lambda_j^* = 0$ then $\gamma_0 = \gamma_1' = 0$, $\beta_{ij} = \beta_{ji} = \theta_j$ and $k^2(\theta_j-\theta_i)(k-\theta_j(2-k))/(2-k(1+\theta_j))^3(2+k(1+\theta_j)) = \gamma_1 = \gamma_0' \geq 0$, by (A.1). As $\theta_j \in [-1,1]$ then $(2-k(1+\theta_j))(2+k(1+\theta_j)) > 0$ and so $(\theta_j-\theta_i)(k-\theta_j(2-k)) \geq 0$.

For uniqueness, we argue that only $\lambda_i^* = 1$, $\lambda_j^* = 0$ solves the Kuhn-Tucker conditions, and as maximum exists in [0,1], it is the unique maximum. We argue by contradiction, letting $(\theta_j-\theta_i)(k-\theta_j(2-k)) \geq 0$ and $\lambda_i^* < 1$ or $\lambda_j^* < 0$ or both. By Lemma 1, neither $\lambda_i = \lambda_j = 1$ nor $\lambda_i = \lambda_j = 0$ are equilibrium profiles.

Case 1: If $\lambda_i^* < 1$ and $\lambda_j^* > 0$: Then $\gamma_1 = \gamma_0' = 0$. If $\theta_i > \theta_j$ then $\theta_j \geq k/(2-k)$ and $\beta_{ij}, \beta_{ji} > k/(2-k)$. If $\theta_j > \theta_i$ then $\theta_j \leq k/(2-k)$ and $\beta_{ij}, \beta_{ji} < k/(2-k)$.

As $4-k^2(1+\beta_{ij})(1+\beta_{ji}) > 0$ for $-1 < k < 1$, both players FOC in (A.1) yield:

$$(\theta_i-\theta_j)k(b_3-\beta_{ij})(\beta_{ji}-b_1) \geq 0 \tag{A.2}$$

$$(\theta_i-\theta_j)k(b_4-\beta_{ij})(\beta_{ji}-b_2) \geq 0 \tag{A.3}$$

with $b_1 = 4\beta_{ij}/(k(2+k-\beta_{ij}(2-k))) - 1$, $b_2 = (1+\beta_{ij})(2+k)k/(4+k(1+\beta_{ij})(2-k))$, $b_3 = (2+k)/(2-k)$, $b_4 = -4/k(2-k) - 1$ and:

$$\frac{(b_2-b_1)(b_3-\beta_{ij})(b_4-\beta_{ij})}{(\beta_{ij}(2-k)-k)} = \frac{4(1+k)(2+k(1+\beta_{ij}))}{k^2(2-k)^2} \geq 0 \tag{A.4}$$

Observe that $b_3 > k/(2-k)$, $b_4 > b_3 \leftrightarrow k < 0$ and $b_4 > k/(2-k) \leftrightarrow k < 0$.

For $k < 0$ and $\theta_i > \theta_j$, then $b_4 > b_3 > k/(2-k)$ and $\beta_{ij}, \beta_{ji} > k/(2-k)$. If $b_3 > \beta_{ij} > k/(2-k)$ then (A.2), (A.3) and (A.4) yield $\beta_{ji} \leq b_1$. But as $\beta_{ij} = b_1$ at $\beta_{ij} = k/(2-k)$ and $\partial b_1/\partial\beta_{ij} = 4(2+k)/k(2+k-\beta_{ij}(2-k))^2 < 0$, then for $\beta_{ij} > k/(2-k)$ we have $\beta_{ji} < k/(2-k)$. A contradiction. If $b_4 > \beta_{ij} > b_3$ then (A.2), (A.3) and (A.4) yield $b_1 \leq \beta_{ji} \leq b_2$ and $b_2 \leq b_1$. A contradiction. If $\beta_{ij} > b_4$ then (A.2), (A.3) and (A.4) yield $\beta_{ji} \geq b_2$. But since $\beta_{ij} = b_2$ at $\beta_{ij} = k/(2-k)$ and $\partial b_2/\partial\beta_{ij} = 4(2+k)k/(4+k(1+\beta_{ij})(2-k))^2 < 0$, then for $\beta_{ij} > k/(2-k)$ we have $\beta_{ji} < k/(2-k)$. A contradiction.

If $k<0$ and $\theta_j>\theta_i$, then $\beta_{ij},\beta_{ji}<k/(2-k)<b_3<b_4$. To wit (A.2), (A.3) and (A.4) dictate $\beta_{ji}\geq b_1$. But as $\beta_{ij}=b_1$ at $\beta_{ij}=k/(2-k)$ and $b_1=\partial b_1/\partial\beta_{ij}<0$, then $\beta_{ij}<k/(2-k)$ yield $\beta_{ji}>k/(2-k)$. A contradiction.

If $0<k<1$ then $b_3>1$ and $b_4<-1$ so $b_2\geq b_1$ iff $\beta_{ij}\leq k/(2-k)$. Now (A.2) and (A.3) dictate $(\theta_j-\theta_i)(\beta_{ji}-b_1)\leq 0$ and $(\theta_j-\theta_i)(\beta_{ji}-b_2)\geq 0$. When $\theta_i>\theta_j$ this reduces to $b_1\leq\beta_{ji}\leq b_2$, and $b_1\geq b_2$, by (A.4). A contradiction. Equivalently, if $\theta_j>\theta_i$ this reduces to $b_2\leq\beta_{ji}\leq b_1$ and $b_2\geq b_1$. A contradiction.

Case 2: If $\lambda_i^*<1$ and $\lambda_j^*=0$: Then $\gamma_1=\gamma_1'=0$, $\beta_{ji}=\theta_j$, $\theta_j<\beta_{ij}\leq\theta_i$ if $\theta_i>\theta_j$ and $\theta_i\leq\beta_{ij}<\theta_j$ if $\theta_j>\theta_i$. In this case the FOC yield (A.2) and the reversed inequality of (A.3). We now use the same logic of the previous case. For $k<0$ and $\theta_i>\theta_j$, if $b_3>\beta_{ij}>k/(2-k)$ then (A.2), (A.3) and (A.4) yield $b_2\leq\beta_{ji}\leq b_1$ and $b_1\leq b_2$. A contradiction. If $b_4>\beta_{ij}>b_3$ then (A.2), (A.3) and (A.4) yield $\beta_{ji}\geq b_1$. But $\partial b_1/\partial\beta_{ij}<0$ and $b_1>1$ at $\beta_{ij}=b_4$, so for $b_4>\beta_{ij}>b_3$ we have $\beta_{ji}>1$. A contradiction. If $\beta_{ij}>b_4$ then (A.2), (A.3) and (A.4) yield $b_1\leq\beta_{ji}\leq b_2$ and $b_2\geq b_1$. But since $\beta_{ij}=b_2$ at $\beta_{ij}=k/(2-k)$ and $\partial b_2/\partial\beta_{ij}<0$, then $\beta_{ij}>k/(2-k)$ yields $\beta_{ji}<k/(2-k)$. A contradiction. For $k<0$ and $\theta_j>\theta_i$, then $\beta_{ij},\beta_{ji}<k/(2-k)<b_3<b_4$. To wit (A.2), (A.3) and (A.4) dictate $b_1\leq\beta_{ji}\leq b_2$ and $b_2\leq b_1$. A contradiction.

If $0<k<1$ then $(\theta_j-\theta_i)(\beta_{ji}-b_1)\leq 0$ and $(\theta_j-\theta_i)(\beta_{ji}-b_2)\leq 0$ by (A.2) and (A.3) and $b_2\leq b_1\leftrightarrow\beta_{ij}\geq k/(2-k)$ by (A.4). When $\theta_i>\theta_j$ then $\beta_{ij}>k/(2-k)$, so the FOC reduce to $\beta_{ji}\geq b_1$. But $b_1-\beta_{ij}=-(2+k(1+\beta_{ij}))(k-\beta_{ij}(2-k))/k(2+k-\beta_{ij}(2-k))$, then $\beta_{ij}\geq b_1\leftrightarrow\beta_{ij}\leq k/(2-k)$. To wit, $\beta_{ij}<b_1\leq\beta_{ji}$. A contradiction. Equivalently, if $\theta_j>\theta_i$ then $\beta_{ij}<k/(2-k)$ so the FOC yield $\beta_{ji}\leq b_1$. To wit, $\beta_{ji}\leq b_1<\beta_{ij}$. A contradiction.

Case 3: If $\lambda_i^*=1$ and $\lambda_j^*>0$: Then $\gamma_0=\gamma_0'=0$, $\beta_{ij}=\theta_j$, $\theta_j<\beta_{ji}\leq\theta_i$ if $\theta_i>\theta_j$ and $\theta_i\leq\beta_{ji}<\theta_j$ if $\theta_j>\theta_i$. In this case the FOC yield the reversed inequality of (A.2) and (A.3). We now use the same logic of the previous case. For $k<0$ and $\theta_i>\theta_j$, if $b_3>\beta_{ij}>k/(2-k)$ then (A.2), (A.3) and (A.4) yield $b_1\leq\beta_{ji}\leq b_2$ and $b_1\leq b_2$. But since $b_2=\beta_{ij}$ at $\beta_{ij}=k/(2-k)$ and $\partial b_2/\partial\beta_{ij}<0$, then $\beta_{ij}>k/(2-k)$ yields $\beta_{ji}<k/(2-k)$. A contradiction. If $b_4>\beta_{ij}>b_3$ then (A.2), (A.3) and (A.4) yield $b_2\leq\beta_{ji}\leq b_1$ and $b_2\leq b_1$. But as $b_1=\beta_{ij}$ at $\beta_{ij}=k/(2-k)$ and $\partial b_1/\partial\beta_{ij}<0$, then for $\beta_{ij}>k/(2-k)$ we have $\beta_{ji}<k/(2-k)$. A contradiction. If $\beta_{ij}>b_4$ then (A.2), (A.3) and (A.4) yield $b_2\leq\beta_{ij}\leq b_1$ and $b_1\geq b_2$. A contradiction.

If $k<0$ and $\theta_j>\theta_i$, then $\beta_{ij},\beta_{ji}<k/(2-k)<b_3<b_4$. To wit (A.2), (A.3) and (A.4) dictate $b_2\leq\beta_{ji}\leq b_1$ and $b_2\leq b_1$. But as $b_1=\beta_{ij}$ at $\beta_{ij}=k/(2-k)$ and $\partial b_2/\partial\beta_{ij}<0$, then for $\beta_{ij}<k/(2-k)$ we have $\beta_{ji}>k/(2-k)$. A contradiction.

If $0<k<1$ then (A.2) and (A.3) dictate $(\theta_j-\theta_i)(\beta_{ji}-b_1)\geq 0$ and $(\theta_j-\theta_i)(\beta_{ji}-b_2)\geq 0$. When $\theta_i>\theta_j$ this reduces to $\beta_{ji}\leq b_2$, as $\beta_{ij}>k/(2-k)$. But $b_2-\beta_{ij}=(2+k(1+\beta_{ij}))(k-\beta_{ij}(2-k))/(4+k(1+\beta_{ij})(2-k))$, then $\beta_{ij}\leq b_2\leftrightarrow\beta_{ij}\leq k/(2-k)$. To wit, $\beta_{ji}\leq b_2<\beta_{ij}$. A contradiction. If $\theta_j>\theta_i$ then $\beta_{ji}\geq b_2$. To wit, $\beta_{ij}<b_2\leq\beta_{ji}$. A contradiction.

For the interior equilibrium, we intersect best responses in (3). This yields two candidates for equilibrium: $\lambda=(\theta_i-\kappa)/(\theta_i-\theta_j)$ and $\lambda'=(2+\theta_i+1/\kappa)/(\theta_i-\theta_j)$. We discard λ' as it is either negative or exceeds one. Letting $\lambda\in(0,1)$ yields $\left(\theta_i-\theta_j\right)\left(\theta_j-\kappa\right)<0$. Behavior is $\beta_{ij}^*=\beta_{ji}^*=\kappa$. □

Proof. Proof of Proposition 2: Integrating (5) by parts yields:

$$\bar{\beta}(\theta_i)=\begin{cases}\theta_i-\int_{\kappa}^{\theta_i}F(\omega)d\omega & \text{for } \theta_i\geq\kappa\\ \kappa-\int_{\theta_i}^{\kappa}F(\omega)d\omega & \text{for } \theta_i\leq\kappa\end{cases}\tag{A.5}$$

We now find the critical values $\underline{\theta}$ and $\bar{\theta}$ such that if $\theta_i\in[\underline{\theta},\bar{\theta}]$ then behavior-reversion might arise. We divide the analysis into cases.

For $\kappa>0$: exploiting (A.5), as $\partial\bar{\beta}(\theta_i)/\partial\theta_i=1-F(\theta_i)\in[0,1)$ for $\theta_i\geq\kappa$, we obtain $0\leq\bar{\beta}(\theta_i)\leq\theta_i$; hence, $\bar{\theta}=\kappa$. If $\theta_i\leq\kappa$ then $\bar{\beta}(\theta_i)\geq\theta_i$ since $\partial\bar{\beta}(\theta_i)/\partial\theta_i=F(\theta_i)\in[0,1)$. If $\bar{\beta}(-1)=\mathbb{E}(\min(\kappa,\Theta_j))<0$ then a unique $\underline{\theta}>-1$ solves $\bar{\beta}(\underline{\theta})=0$ and so $\theta_i\leq\bar{\beta}(\theta_i)\leq 0$ for $\theta_i\leq\underline{\theta}$. Easily, as $\bar{\beta}(0)>0$, then $\underline{\theta}<0$. If $\bar{\beta}(-1)\geq 0$ then $\underline{\theta}=-1$. To wit, $\bar{\beta}(\theta_i)\geq\max(0,\theta_i)$ if $\underline{\theta}\leq\theta_i\leq\bar{\theta}$.

For $\kappa<0$: if $\theta_i\leq\kappa$ then $\theta_i\leq\bar{\beta}(\theta_i)\leq 0$ and so $\underline{\theta}=\kappa$, by (A.5). Next, for $\theta_i\geq\kappa$ we have $\partial\bar{\beta}(\theta_i)/\partial\theta_i=1-F(\theta_i)\in[0,1)$ and $\bar{\beta}(0)<0$, by (A.5). To wit, a unique $\bar{\theta}>0$ solves $\bar{\beta}(\bar{\theta})=0$ and for $\theta_i\geq\bar{\theta}$ then $0\leq\bar{\beta}(\theta_i)\leq\theta_i$. We have $\bar{\theta}<1$ iff $\bar{\beta}(1)=\mathbb{E}(\max(\kappa,\Theta_j))>0$. Easily, $\bar{\beta}(\theta_i)\leq\min(0,\theta_i)$ if $\underline{\theta}\leq\theta_i\leq\bar{\theta}$. □

Proof. Proof of Proposition 3: We optimize (9) in λ_i. Observe that given our linear specification for preferences, $\partial b_i(\theta_i)/\partial\lambda_i=\mu_j-\theta_i$. Fix $\bar{b}_j$ and $\bar{b}_i$; then, computing the FOC we get:

$$\frac{\partial\mathbb{E}_{\theta_j}[\Pi_i]}{\partial\lambda_i}=\frac{-k^2(2+k(1+\bar{b}_j))^2b_i^*(\theta_i)(\bar{\theta}_j-\theta_i)}{2(4-k^2(1+\bar{b}_i)(1+\bar{b}_j))^2}=0\leftrightarrow b_i^*(\theta_i)=0$$

Clearly, if $\theta_i=\mu$ then any λ_i is optimal. Otherwise, the solution is $\lambda_i^*=\theta_i/(\theta_i-\mu)$. To guarantee $0\leq\lambda_i^*\leq 1$ we restrict types to $\max(\theta_i,\mu)\geq 0\geq\min(\theta_i,\mu)$. Otherwise, if $\theta_i>\mu\geq 0$ or $0\geq\mu>\theta_i$ then $\lambda_i^*=1$, and $\lambda_i^*=0$ if $0\geq\theta_i>\mu$ or $\mu>\theta_i\geq 0$. □

Proof. Proof of Proposition 4: As (10) is analogous to (4) but when $\kappa=0$ and $\theta_j=\mu$ we deduce that the expected preferences in this case are as in (5). As $\bar{\beta}(\theta_i)$ rises in θ_i and obeys $\bar{\beta}(\theta_i)=\theta_i$ only at $\theta_i=\kappa$, it follows then that when $\kappa=0$ we have $\theta_i>0$ if and only if $\bar{\beta}(\theta_i)>0$. That is, there is no behavior reversion. □

Lemma 1. *For $k\neq 0$, neither $\lambda_i=\lambda_j=1$ nor $\lambda_i=\lambda_j=0$ are equilibrium profiles.*

Proof. We show that no $\theta_i,\theta_j\in[-1,1]$ solve simultaneously the FOC in (A.1), which are necessary for a maximum. In either case, the FOC dictate:

$$(\theta_i-\theta_j)(4\theta_j-k(1+\theta_i)(2-k)(b_3-\theta_j))\geq 0\tag{A.6}$$

$$(\theta_j - \theta_i)(4\theta_i - k(1+\theta_j)(2-k)(b_3-\theta_i)) \geq 0 \tag{A.7}$$

If $\theta_i > \theta_j$ then $k(1+\theta_i)(2-k)(b_3-\theta_j) \leq 4\theta_j < 4\theta_i \leq k(1+\theta_j)(2-k)(b_3-\theta_i)$ and so $k(2-k)(\theta_i-\theta_j)(b_3+1) < 0$; a contradiction for $k > 0$. The $\theta_j > \theta_i$ case is analogous. If $k < 0$ and $\theta_i > \theta_j$ then $\theta_i \in [0, b_3]$ clearly does not solve (A.7). Neither does $\theta_i \in (b_3, 1]$, as $k(1+\theta_j)(2-k)(b_3-\theta_i)$ rises linearly from 0 to $2k^2(1+\theta_j) < 4k^2 < 4$. The only candidates are $\theta_j < \theta_i < 0$. But in this case the FOC yield:

$$\frac{k(1+\theta_i)(2+k)}{4+k(1+\theta_i)(2-k)} \leq \theta_j \leq \frac{4\theta_i}{k(2-k)(b_3-\theta_i)} - 1 \tag{A.8}$$

This inequality limits are equal at $\theta_i = k/(2-k)$. Easily, their slopes are negative, and the upper limit slope exceeds the lower limit slope iff $-2 - 2\theta_i k + k^2 + \theta_i k^2 \geq 0$, which is iff $-k(2-k)(\theta_i + (2-k^2)/k(2-k)) \geq 0$. A contradiction. To wit, the interval in (A.8) is empty. The $\theta_j > \theta_i$ case is analogous. □

A2. Public good contribution game with strategic reciprocity

In this section, we solve the public good contribution game (Levine, 1998) accounting for endogenous reciprocity. In the short-run each player i must independently choose how much to contribute to the provision of a public good (i.e., number of tokens x_i) to maximize perceived utilities $u_i(x_i, x_j) = \pi_i(x_i, x_j) + \beta_{ij}\pi_j(x_j, x_i)$, where material payoffs are $\pi_i(x_i, x_j) = -x_i + \gamma(x_i + x_j)$ with $0 < \gamma < 1$. Due to the linearity of the perceived utility, the first order condition for cooperation is $-1 + \gamma + \beta_{ij}\gamma \geq 0$. That is, player i cooperates only if $\beta_{ij} \geq (1-\gamma)/\gamma \equiv \phi$. Normalizing the total number of available tokens per player to 1, we then obtain

$$x_i^* = \begin{cases} 1 & \text{if } \beta_{ij} \geq \phi \\ 0 & \text{if } \beta_{ij} \leq \phi \end{cases}$$

Next, given our interdependent preference structure, similar to Levine (1998), player i cooperates only if $\theta_i + \lambda_i(\theta_j - \theta_i) \geq \phi$. To wit,

$$x_i^* = 1 \leftrightarrow \beta_{ij} \geq \phi \leftrightarrow \theta_i \geq \frac{\phi - \lambda_i\theta_j}{1-\lambda_i}$$

That is, as in Levine (1998), only players sufficiently altruistic decide to cooperate. However, and unlike them, in the long-run players are able to choose how much to reciprocate; that is, the specific value of λ_i. To solve for this optimal decision we write player i long-run payoffs

$$\Pi_i = \pi_i(x_i^*, x_j^*) = -x_i^* + \gamma(x_i^* + x_j^*) = x_i^*(\gamma - 1) + \gamma x_j^*$$

As $0 < \gamma < 1$, the long-run payoffs fall in the optimal choice x_i^*. This means that player i would like to choose λ_i in such a way that induces $\beta_{ij} \leq \phi$ and thus $x_i^* = 0$. Formally, the optimal reciprocity choice is then:

$$\lambda_i^* \in \begin{cases} [0,1] & \text{if } \min(\theta_i, \theta_j) \geq \phi \\ \left[0, \frac{\phi-\theta_i}{\theta_j-\theta_i}\right] & \text{if } \theta_j \geq \phi \geq \theta_i \\ [0,1] & \text{if } \phi \geq \max(\theta_i, \theta_j) \\ \left[\frac{\theta_i-\phi}{\theta_i-\theta_j}, 1\right] & \text{if } \theta_i \geq \phi \geq \theta_j \end{cases}$$

Only in the first case, when $\min(\theta_i, \theta_j) \geq \phi$, we obtain that in equilibrium $\beta_{ij}^* \geq \phi$ and also $\beta_{ji}^* \geq \phi$, hence $x_i^* = x_j^* = 1$. That is, only when players are both sufficiently altruistic we obtain that they both endogenously induce themselves to cooperate, which is consistent with Levine (1998) finding. However, in any other case there is no cooperation as players are able to induce $\beta_{ij}^* \leq \phi$ and $\beta_{ji}^* \leq \phi$ by adjusting how much they reciprocate. This insight radically differs from Levine (1998) prediction. It states that if players can choose how much to reciprocate they will both try to behave sufficiently spiteful; unless, of course, they are intrinsically restricted to do so.

References

Alger, I., Weibull, J.W., 2013. Homo Moralis–preference evolution under incomplete information and assortative matching. Econometrica 81 (6), 2269–2302.
Alger, I., Weibull, J.W., 2019. Evolutionary models of preference formation. Annu. Rev. Econ. 11 (1), 329–354. doi:10.1146/annurev-economics.
Bester, H., Güth, W., 1998. Is altruism evolutionarily stable? J. Econ. Behav. Organ. 34 (2), 193–209. doi:10.1016/S0167-2681(97)00060-7.
Bolle, F., 2000. Is altruism evolutionarily stable? And envy and malevolence?: Remarks on Bester and Güth. J. Econ. Behav. Organ. 42 (1), 131–133.
Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. Am. Econ. Rev. 90 (1), 166–193. http://www.jstor.org/stable/117286.
Carrasco, J.A., Harrison, R., Villena, M., 2018. Interdependent preferences and endogenous reciprocity. J. Behav. Exp. Econ. 76, 68–75. doi:10.1016/j.socec.2018.08.002.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117 (3), 817–869. http://www.jstor.org/stable/4132490.
Coleman, J.S., 1990. Foundations of Social Theory/James S. Coleman. Belknap Press of Harvard University Press Cambridge, Mass.
Dekel, E., Ely, J.C., Yilankaya, O., 2007. Evolution of preferences. Rev. Econ. Stud. 74 (3), 685–704. http://www.jstor.org/stable/4626157.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47 (2), 268–298. doi:10.1016/j.geb.2003.06.003.
Dufwenberg, M., Kirchsteiger, G., 2019. Modelling kindness. J. Econ. Behav. Organ. 167, 228–234.

Ely, J.C., Yilankaya, O., 2001. Nash equilibrium and the evolution of preferences. J. Econ. Theory 97 (2), 255–272.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54 (2), 293–315.
Fehr, E., Fischbacher, U., Gächter, S., 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. Hum. Nat. 13 (1), 1–25.
Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does fairness prevent market clearing? An experimental investigation. Q. J. Econ. 108 (2), 437–459.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114 (3), 817–868. http://www.jstor.org/stable/2586885.
Fershtman, C., Segal, U., 2018. Preferences and social influence. Am. Econ. J. Microecon. 10 (3), 124–142. doi:10.1257/mic.20160190.
Güth, W., Napel, S., 2006. Inequality aversion in a variety of games – an indirect evolutionary analysis*. Econ. J. 116 (514), 1037–1056. doi:10.1111/j.1468-0297.2006.01122.x.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3 (4), 367–388.
Heifetz, A., Shannon, C., Spiegel, Y., 2007a. The dynamic evolution of preferences. Econ. Theory 32 (2), 251–286. http://www.jstor.org/stable/27822555
Heifetz, A., Shannon, C., Spiegel, Y., 2007b. What to maximize if you must. J. Econ. Theory 133 (1), 31–57.
Isaac, R.M., Walker, J.M., 1988. Group size effects in public goods provision: the voluntary contributions mechanism. Q. J. Econ. 103 (1), 179–199.
Isoni, A., Sugden, R., 2019. Reciprocity and the paradox of trust in psychological game theory. J. Econ. Behav. Organ. 167, 219–227.
Koçkesen, L., Ok, E.A., Sethi, R., 2000. The strategic advantage of negatively interdependent preferences. J. Econ. Theory 92 (2), 274–299. doi:10.1006/jeth.1999.2587.
Levine, D.K., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dyn. 1 (3), 593–622. doi:10.1006/redy.1998.0023.
McCabe, K., Smith, V., 2000. Goodwill accounting in economic exchange. In: Bounded Rationality: The Adaptive Toolbox. MIT Press, Cambridge, MA, pp. 319–340.
Ok, E., Vega-Redondo, F., 2001. On the evolution of individualistic preferences: an incomplete information scenario. J. Econ. Theory 97 (2), 231–254.
Okuno-Fujiwara, M., Postlewaite, A., 1995. Social norms and random matching games. Games Econ. Behav. 9 (1), 79–109.
Possajennikov, A., 2000. On the evolutionary stability of altruistic and spiteful preferences. J. Econ. Behav. Organ. 42 (1), 125–129. doi:10.1016/S0167-2681(00)00078-0.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83 (5), 1281–1302.
Rotemberg, J.J., 1994. Human relations in the workplace. J. Polit. Economy 102 (4), 684–717. doi:10.1086/261951.
Sethi, R., Somanathan, E., 2001. Preference evolution and reciprocity. J. Econ. Theory 97 (2), 273–297.
Sobel, J., 2005. Interdependent preferences and reciprocity. J. Econ. Lit. 43 (2), 392–436.
Villena, M.G., Zecchetto, F., 2011. Subject-specific performance information can worsen the tragedy of the commons: experimental evidence. J. Econ. Psychol. 32 (3), 330–347.